Title: Unravelling genome-wide mosaic microsatellite mutations at single-cell resolution

Authors:

Chunyi Wang[1,2,8], Wenxuan Fan[1,2,8], Weixiang Wang[1,2,3,8], Yonghe Xia[1,2,8], Jinhong Lu[1,2,3,8], Xiaoyu Ma[1,2,8], Jichuan Yu[1,2,8], Yunchao Zheng[1,2], Qing Yang[1,2,4], Meizhen Lin[1,2], Xiuyan Yu[5], Xushen Xiong[6], Manolis Kellis[7], Qi Xie[1,2*], Yanmei Dou[1,2*], Data Group

1. School of Life Sciences, Westlake University, Hangzhou, Zhejiang, 310024, China
2. Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, 310024, China
3. Fudan University, Shanghai, 200433, China
4. College of Life Sciences, Zhejiang University, Hangzhou, Zhejiang, China
5. Department of Breast Surgery, Second Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang, China
6. Liangzhu Laboratory, Zhejiang University, Hangzhou, Zhejiang, China
7. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
8. These authors contributed equally.

*Correspondence: xieqi@westlake.edu.cn (Q.X.), douyanmei@westlake.edu.cn (Y.D.).

Abstract:

Short tandem repeats (STRs), or microsatellites, are highly mutable genomic elements that modulate gene regulations and are implicated in a range of human diseases. However, detecting mosaic STR mutations at single-cell resolution remains challenging due to both technical and biological complexities. To address this, we developed BayesMonSTR, a robust algorithm that enables accurate detection of mosaic STR mutations. Using this tool in single-cell analysis of human tissues, we reveal an accumulation of longer mosaic STR insertions and deletions (indels) in aging mitotic and post-mitotic cells. Strikingly, prefrontal cortex (PFC) neurons accumulate a higher burden of STR mutations than B cells or lung epithelium, with aged neurons exhibiting a particularly pronounced increase in longer STR deletions. These mutations are enriched at transcription start sites (TSSs) and active enhancers of highly expressed genes. Our work establishes a foundation for genome-wide, hypothesis-free discovery of disease-associated mosaic STR mutations and reveals a previously unexplored landscape of mosaic STR variation in development and aging.

Somatic mosaicism, arising from postzygotic mutations accumulated throughout life, represents the dynamic interface between genome instability, environmental exposures, and cellular repair processes, and is a recognized contributor to aging and disease[1]. STR regions are vastly overrepresented in the human genome[2], and their intrinsic hypermutability—occurring at a rate roughly 100 times that of non-repeat regions[3]—renders them particularly vulnerable to mutations. Although STRs play essential roles in regulating transcription, facilitating chromatin interactions, and are linked to nearly a hundred human disorders[4,5], mosaic STR mutations in aging tissues remain largely unexplored. Current genotyping methods[6–8] are unable to resolve mosaic STR mutations at single-cell resolution, constrained by both technical challenges (e.g., amplification & sequence errors, alignment ambiguities) and the inherent biological complexity of repetitive regions (e.g., repeat sequence interruptions, truncated motifs).

Recognizing this critical gap, we developed BayesMonSTR—a computational framework that integrates probabilistic read segmentation, haplotype phasing, a hierarchical Bayesian network, and a machine learning classification model to detect mosaic STR insertions, deletions (indels) and mismatches from single cells with nucleotide-level resolution (Fig. 1a-f, Supplementary Fig. 1, Supplementary Table 1, Methods). By integrating multiple layers of data—at the population, individual, and single-cell levels—and probabilistically quantifying all aspects (e.g., sequence error rate, read phasing probability, repeat segmentation, and genotyping likelihoods), BayesMonSTR resolves ambiguities in both extracting imperfect STR alleles from noisy data and robust detection of mutation candidates. Furthermore, it rigorously filters out technical artifacts and germline polymorphisms using a machine learning model trained based on a multifaceted set of "gold-standard" orthogonal benchmarking data generated in this study (Fig. 1g-h; see Supplementary Information for details on the validation strategies: duplicate single-cell libraries facilitated by single-cell DNA & RNA cross validation, and deep bulk DNA sequencing from adjacent tissues).

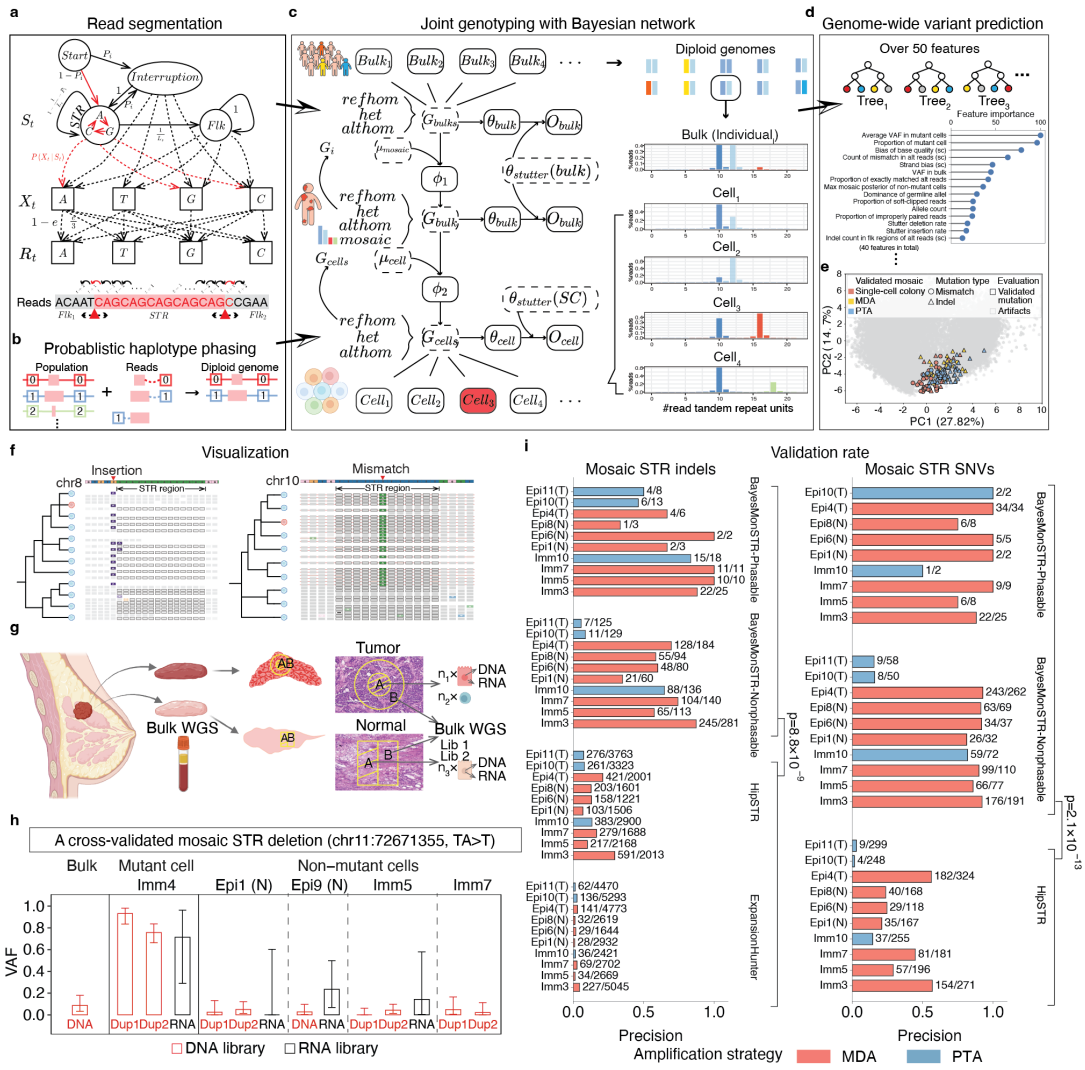


Fig. 1 Algorithm and benchmark of BayesMonSTR. (**a**) BayesMonSTR employs an HMM to probabilistically segment each sequencing read, distinguishing the STR region from its flanking sequences. (**b**) STR alleles are assigned to their chromosomal haplotypes by statistically linking them to the alleles of neighboring heterozygous variants. (**c**) A Hierarchical Bayesian Network integrates three layers of data—population-level sequencing, individual-level bulk sequencing, and single cell sequencing—to calculate individual- and cell-level genotype posteriors. (**d**) A random forest model, trained on over 50 carefully chosen read-level and genome-level features, classifies true mosaic mutations from artifacts. (**e**) PCA constructed from top 40 features shows that validated mosaic variants from independent single-cell datasets cluster in similar regions, highlighting the robustness of BayesMonSTR. (**f**) Illustration of two mosaic STR mutations from real single cell data. A 1-bp insertion detected in single-cell colony data is shown on the left. Supporting sequencing reads covering the variant site are displayed for the cell harboring the mutation. A base substitution within STR region, identified from single-cell whole-genome sequencing (scWGS) data is shown on the right. The plots were generated with BayesMonSTR-INSIGHT[9]. (**g**) A gold-standard benchmarking dataset was generated from a breast tissue sample,

including tumor epithelial cells, normal epithelial cells and immune cells. Duplicate single-cell libraries, RNA-seq data from the same single cell, and bulk sequencing data from adjacent tissues were generated for cross validation. cross validation, and deep bulk DNA sequencing from adjacent tissues. (**h**) Example variant validated by duplicate library sequencing and RNA-seq from the same cell, with mutant reads also present in adjacent-tissue bulk DNA-seq. Error bars: 95% CI of mutant-cell VAFs (binomial sampling). (**i**) BayesMonSTR achieved multi-fold higher accuracy in detecting cell-specific mosaic STR mutations compared to adapted HipSTR and ExpansionHunter (two-tailed t-test).

For the ground-truth benchmarking, we sequenced 29 single cells from a breast tissue sample, including seven tumor epithelial cells, 12 normal epithelial cells, and ten immune cells (Methods, Supplementary Fig. 2, Supplementary Table 2). Single-cell DNA was amplified using two widely adopted strategies: MDA and PTA (Methods). Across cells, BayesMonSTR consistently achieved ~60–80% validation rates due to its comprehensive framework—several folds higher than HipSTR(v0.6.2)[6] and ExpansionHunter(v5.0.0)[8]($p = 8.8 \times 10^{-9}$ for indels and $p = 2.1 \times 10^{-13}$ for mismatches, two-tailed t-tests, Fig. 1i, Supplementary Fig. 3), two state-of-the-art germline STR mutation callers that we adapted for mosaic mutation detection by applying a series of hard filters (Methods). Of note, precision and recall were lower in some PTA-amplified cells, likely due to additional PCR cycles (9–10 rounds) that introduced DNA-slippage errors (Supplementary Fig. 4). This superior precision was consistently observed across additional independent benchmarks[10,11]—including public MDA-amplified cells, public single-cell colonies and simulated data—while retaining competitive sensitivity (Methods, Supplementary Tables 3-5, Supplementary Figs. 5-6).

Given the hypermutable nature of STR regions, we further investigated their potential for in vivo lineage tracing. Notably, a phylogeny inferred from shared STR mutations in our benchmark dataset—spanning ~1% of the genome—roughly recapitulated the topology of a tree constructed from mosaic single-nucleotide variants (SNVs) identified in the remaining 99% of the genome (Methods, Supplementary Fig. 5h, Supplementary Table 6). These results support the use of STR mutations as a cost-effective approach for lineage tracing.

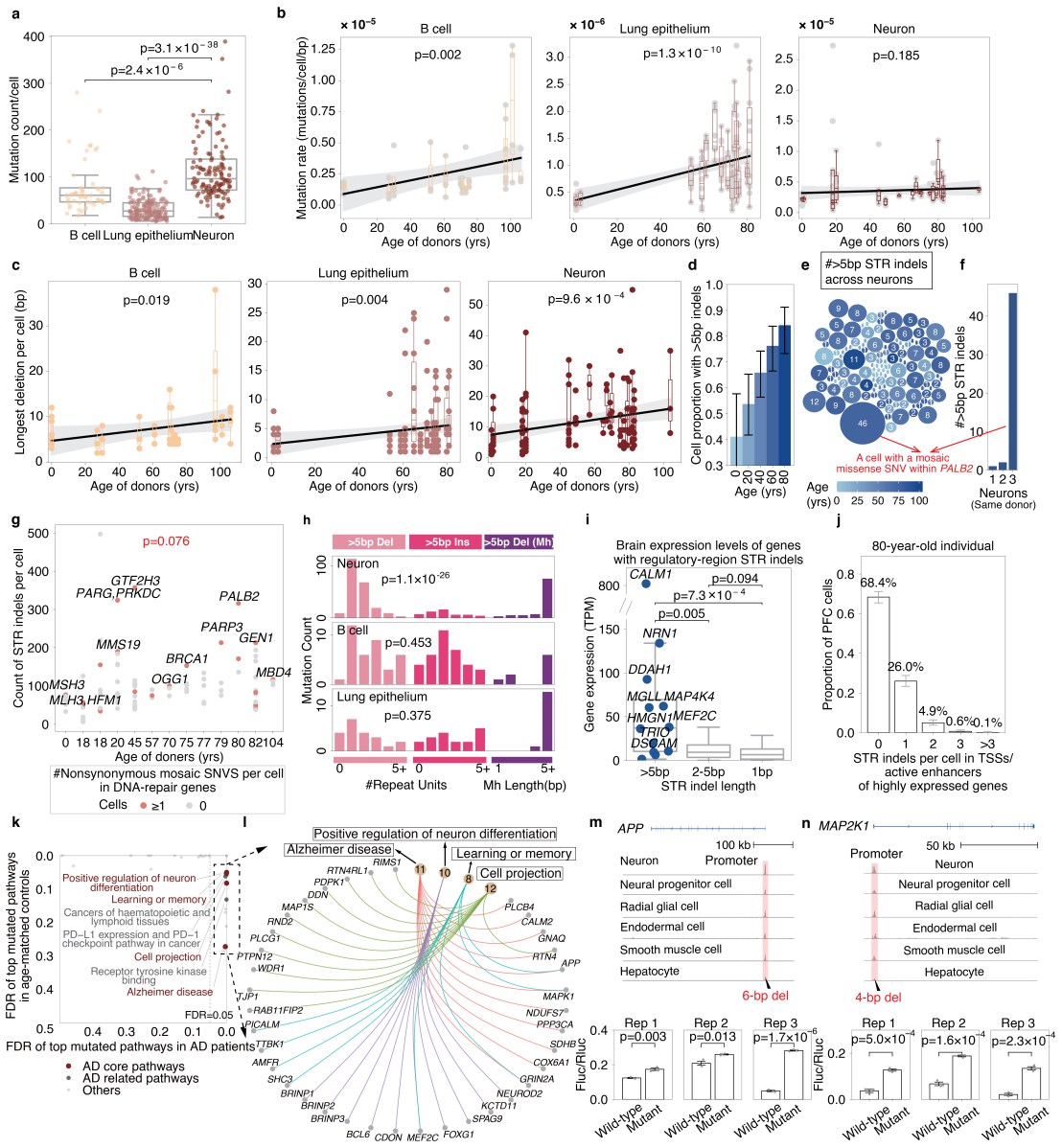


Fig. 2 Mosaic STR mutations across tissues in aging and disease. (**a**) Mosaic STR mutation burden (including indels and SNVs) per cell is significantly higher in neurons than in other cell types (two-tailed Mann-Whitney U test). (**b**) Mosaic mutations within and flanking STR regions accumulate with age in B cells and lung epithelial cells but remain stable in non-dividing neurons (one-tailed t-test for slope >0). Shading indicates 95% CIs. Longer STR indels with high signal-to-noise ratios were used in panels c-j and passed manual validation. (**c**) The longest mosaic STR contraction per cell increases significantly with age across all cell types (same regression approach as in b). (**d**) Estimated proportion of neurons carrying mosaic STR indels >5 bp increases with age (binomial Generalized Linear Model (GLM)). Error bars indicate 95% confidence intervals derived from the fitted model. All insertions >5 bp in panels d-g were manually validated. (**e**) STR indels >5 bp are prevalent in neurons, especially with age. Circle size indicates indel count per neuron, color denotes donor age. The neuron with 46 indels carries a missense mutation in DNA repair gene *PALB2*. (**f**) Distribution of >5

130 bp STR indels in three neurons from the same individual. The *PALB2*-mutant neuron
131 shows substantially higher burden. (**g**) Nonsynonymous mosaic SNVs in DNA repair
132 genes are marginally associated with increased STR indel counts in neurons
133 (Generalized Linear Mixed Models (GLMM) with negative binomial distribution,
134 random intercept for individual). Red points indicate cells with repair-gene SNVs. (**h**)
135 Mutation signatures of >5 bp STR indels across cell types. Mh, microhomology. (**i**)
136 Genes with >5 bp STR indels in TSSs or active enhancers show higher expression than
137 genes with shorter indels (two-tailed Mann-Whitney U test). (**j**) Poisson modeling
138 estimates that >30% of neurons in 80-year-olds carry at least one mosaic STR indel in
139 regulatory regions of highly expressed genes. Error bars indicate 95% confidence
140 intervals computed using binomial Wilson score intervals. (**k**) In AD patients only, STR
141 indels in TSSs and active enhancers are significantly enriched in AD-related genes
142 (Methods). (**l**) Key AD genes contain multi-hit STR indels in regulatory regions across
143 patients. "earlyAD," "lateAD" indicate disease stage. (**m,n**) Luciferase assays in
144 HEK293T cells show that a 6-bp deletion in the *APP* TSS and a 4-bp deletion in the
145 MAP2K1 TSS significantly enhance gene expression (two-tailed t-test, n=3 biological
146 replicates, error bars indicate SD). Normalized H3K4me3 and H3K27ac levels from
147 ENCODE are shown; pink regions indicate promoter/enhancer regions, black triangles
148 indicate indels.

149

150 Using this tool, we characterized the landscape of mosaic STR mutations in three
151 mitotic and post-mitotic tissues, including B cells from blood[12] (56 cells from 14
152 healthy donors aged 0–106 years), lung epithelial cells[10] (160 cells from 16 healthy
153 donors aged 1–81 years) and neurons[13] (126 cells from 19 healthy donors aged 0–104
154 years) (Supplementary Table 7). The mutation rate per cell per base pair ranges from
155 approximately 10^{-6} to 10^{-5}, significantly higher than the reported point mutation rate
156 outside of microsatellite regions[11]. The overall exonic indel mutation rate was about 60%
157 lower than expected ($p = 5.5\times10^{-122}$, one-tailed binomial test, Supplementary Table 8).
158 This reduction was not observed for exonic SNVs (p=0.996, one-tailed binomial test),
159 suggesting strong purifying selection specifically targeting exonic STR indels.

160 Strikingly, neurons—which exit the cell cycle early in development—harbored
161 significantly more mosaic STR mutations than mitotic B cells ($p = 2.4 \times 10^{-6}$) and lung
162 epithelium ($p = 3.1 \times 10^{-38}$, two-tailed Mann-Whitney U tests, Fig. 2a-b). Size-resolved
163 analysis of indel mutation rates revealed an intriguing dichotomy in age-related
164 dynamics. While B cells ($p = 0.003$) and lung epithelium ($p = 1.8 \times 10^{-10}$) accumulated
165 significantly more 1-bp indels—the most abundant indel type, accounting for 79% of
166 all STR indels—with age (one-tailed t-tests), neuronal 1-bp indel burden showed no
167 age-related increase ($p = 0.352$, one-tailed t-test, Supplementary Fig. 7a), indicating
168 that 1-bp STR indels were largely acquired during early development in neurons. By
169 contrast, all three tissues, particularly neurons, accumulated longer indels during aging
170 (Fig. 2c, Supplementary Fig. 7b-d). These findings suggest that longer STR indels arise

171 via distinct, age-dependent mechanisms that operate across cell types, whereas the
172 accumulation of 1-bp indels is primarily linked to replicative history. Beyond indels,
173 all three tissues also accumulated significantly more mosaic SNVs with age
174 (Supplementary Fig. 8).

175 Longer indels affect more genomic sequences and thus pose a greater risk of disrupting
176 DNA structure. Among the three tissues, long STR indels (>5 bp) were significantly
177 more prevalent in neurons than in B cells or lung epithelium ($p = 2.8 \times 10^{-3}$ and $p = 1.0$
178 $\times 10^{-6}$, respectively, two tailed t-tests). By age 80, approximately 84.3% (95% CI:
179 73.3–91.2%) of neurons from healthy individuals carried at least one such indel (Fig.
180 2d, Methods). Across 126 neurons from 19 donors, most cells harbored fewer than ten
181 long indels, though a subset showed markedly higher burdens (Fig. 2e). Notably, one
182 neuron from an 80-year-old female exhibited nearly 50 long STR indels—a >30-fold
183 excess over two other neurons from the same individual (Fig. 2f). This same cell also
184 carried a nonsynonymous mosaic exonic SNV in *PALB2* (Supplementary Table 9), a
185 gene essential for DNA repair. More broadly, neurons harboring nonsynonymous SNVs
186 in DNA repair genes showed a marginally significant increase in STR indel burden (p
187 $= 0.076$, negative binomial GLMM, Fig. 2g, Methods), suggesting that mosaic
188 mutations in DNA repair genes may compromise microsatellite stability during aging.

189 Notably, among >5bp mosaic indels lacking microhomology at mutation
190 boundaries, >5 bp insertions and deletions occurred at similar frequencies in B cells (p
191 $= 0.453$) and lung epithelium ($p = 0.375$), whereas neurons displayed a significant bias
192 toward deletions ($p = 1.1 \times 10^{-26}$, two-tailed binomial tests, Fig. 2h), suggesting a
193 specialized long-deletion mechanism in aging neurons. Concurrently, all three tissues
194 accumulated a substantial proportion of >5bp deletions flanked microhomology—a
195 hallmark of microhomology-mediated end joining (MMEJ), which typically generates
196 long deletions. These accounted for 25.9% of all >5 bp indels in neurons, compared to
197 11.1% in B cells and 26.4% in lung epithelium, suggesting increased double-strand
198 break formation and preferential MMEJ repair in aging tissues[14].

199 Beyond indels, mosaic SNVs within STR regions largely mirrored established extra-
200 STR patterns attributed to oxidative damage and cytosine deamination (Supplementary
201 Fig. 9). Together with our indel findings, these results reveal a complex landscape of
202 microsatellite mutagenesis in aging, characterized by both genomic instability and
203 engagement of specific DNA repair processes.

204 Given the role of microsatellites in gene regulation, we examined STR indel distribution
205 in neuronal regulatory regions. These mutations were enriched in TSSs and active
206 enhancers of genes highly expressed in the brain ($p = 5.0 \times 10^{-3}$, one-tailed t-test,
207 Supplementary Fig. 10a), with longer indels showing the strongest enrichment (Fig. 2i).
208 Based on Poisson sampling, over 30% of neurons in an 80-year-old would harbor at
209 least one such indel (Fig. 2j).

210 To further explore this, we adapted BayesMonSTR for snATAC-seq data (Methods).
211 Analysis of prefrontal cortex from 55 AD patients and 55 controls[15,16] revealed

212 mutation rates comparable to scWGS data (Supplementary Fig. 10b), with a similar
213 excess of long deletions ($p = 1.1 \times 10^{-31}$, one-tailed binomial test, Supplementary Fig.
214 10c). STR indels in AD patients were significantly enriched in TSSs and active
215 enhancers of AD-related genes and pathways (Fig. 2k-l, Supplementary Fig. 10d),
216 including well-established risk genes such as *APP, CD2AP, PICALM*, and *MEF2C*
217 (Supplementary Table 10). Luciferase assays of six STR indels in TSS/enhancer
218 regions of AD-related genes revealed that four significantly disrupted gene expression
219 in 293T cells (Fig. 2m-n, Supplementary Fig. 11), indicating that microsatellites in
220 active regulatory regions are mutation-prone and can affect expression.

221 As the first tool for genome-wide detection of somatic STR mutations in single cells,
222 BayesMonSTR enables hypothesis-free discovery of STR mutagenesis in aging and
223 disease. It also provides a unified toolkit for scWGS, snATAC-seq (BayesMonSTR-
224 ATAC), and bulk WGS (BayesMonSTR-BulkMonSTR)[17], as well as enhanced
225 visualization of STR mutations (BayesMonSTR-INSIGHT)[9], opening new
226 opportunities to investigate the contribution of STR instability to aging and disease.
227 BayesMonSTR is implemented in Python and R and is licensed under the MIT License.
228 The source code, documentation and examples are available on GitHub at
229 https://github.com/douymLab/BayesMonSTR.

230 **Data Group**

231 Chengyu Li[6], Huan Liu[1,2], Yan Luo[1,2], Wenlong Li[1,2], Yangning Lan[1,2], Xiaodong
232 Liu[1,2], Shang Cai[1,2], Danyang He[1,2], Dan Zhou[9]
233 9. The Second Affiliated Hospital, School of Public Health, Zhejiang University
234 School of Medicine, Hangzhou, China

235 **References**

236 1. Shao, D. D. *et al.* Advances in single-cell DNA sequencing enable insights into
237 human somatic mosaicism. *Nat. Rev. Genet.* **26**, 761–774 (2025).

238 2. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev.*
239 *Genet.* **5**, 435–445 (2004).

240 3. Porubsky, D. *et al.* Human de novo mutation rates from a four-generation pedigree
241 reference. *Nature* **643**, 427–436 (2025).

242 4. Rajan-Babu, I.-S., Dolzhenko, E., Eberle, M. A. & Friedman, J. M. Sequence
243 composition changes in short tandem repeats: heterogeneity, detection, mechanisms
244 and clinical implications. *Nat. Rev. Genet.* 1–24 (2024) doi:10.1038/s41576-024-
245 00696-z.

246 5. Horton, C. A. *et al.* Short tandem repeats bind transcription factors to tune eukaryotic
247 gene expression. *Science* **381**, eadd1250 (2023).

6. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat Methods* **14**, 590–592 (2017).

7. Sehgal, A., Ziaei-Jam, H., Shen, A. & Gymrek, M. Genome-wide detection of somatic mosaicism at short tandem repeats. *Bioinformatics* btae485 (2024) doi:10.1093/bioinformatics/btae485.

8. Dolzhenko, E. *et al.* ExpansionHunter: A sequence-graph based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).

9. Xia, Y. & Dou, Y. INSIGHT: A genome data visualization toolkit. *https://github.com/douymLab/INSIGHT* (2026).

10. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).

11. Luquette, L. J. *et al.* Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat Genet* **54**, 1564–1571 (2022).

12. Zhang, L. *et al.* Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci.* **116**, 9014–9019 (2019).

13. A., L., Michael *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).

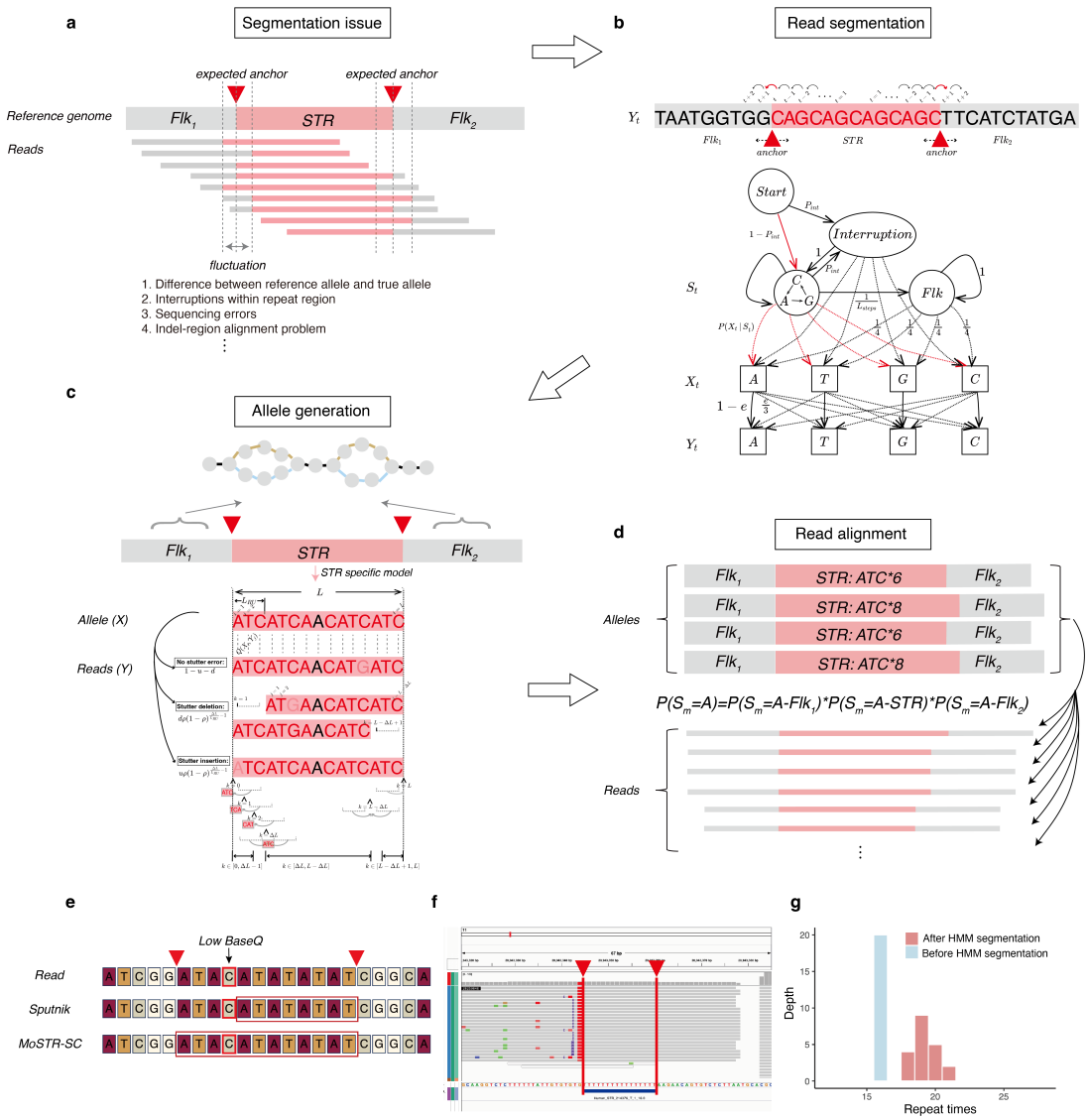
14. Sfeir, A. & Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem. Sci.* **40**, 701–714 (2015).

15. Xiong, X. *et al.* Epigenomic dissection of Alzheimer's disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422-4437.e21 (2023).

16. Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet.* **53**, 1143–1155 (2021).

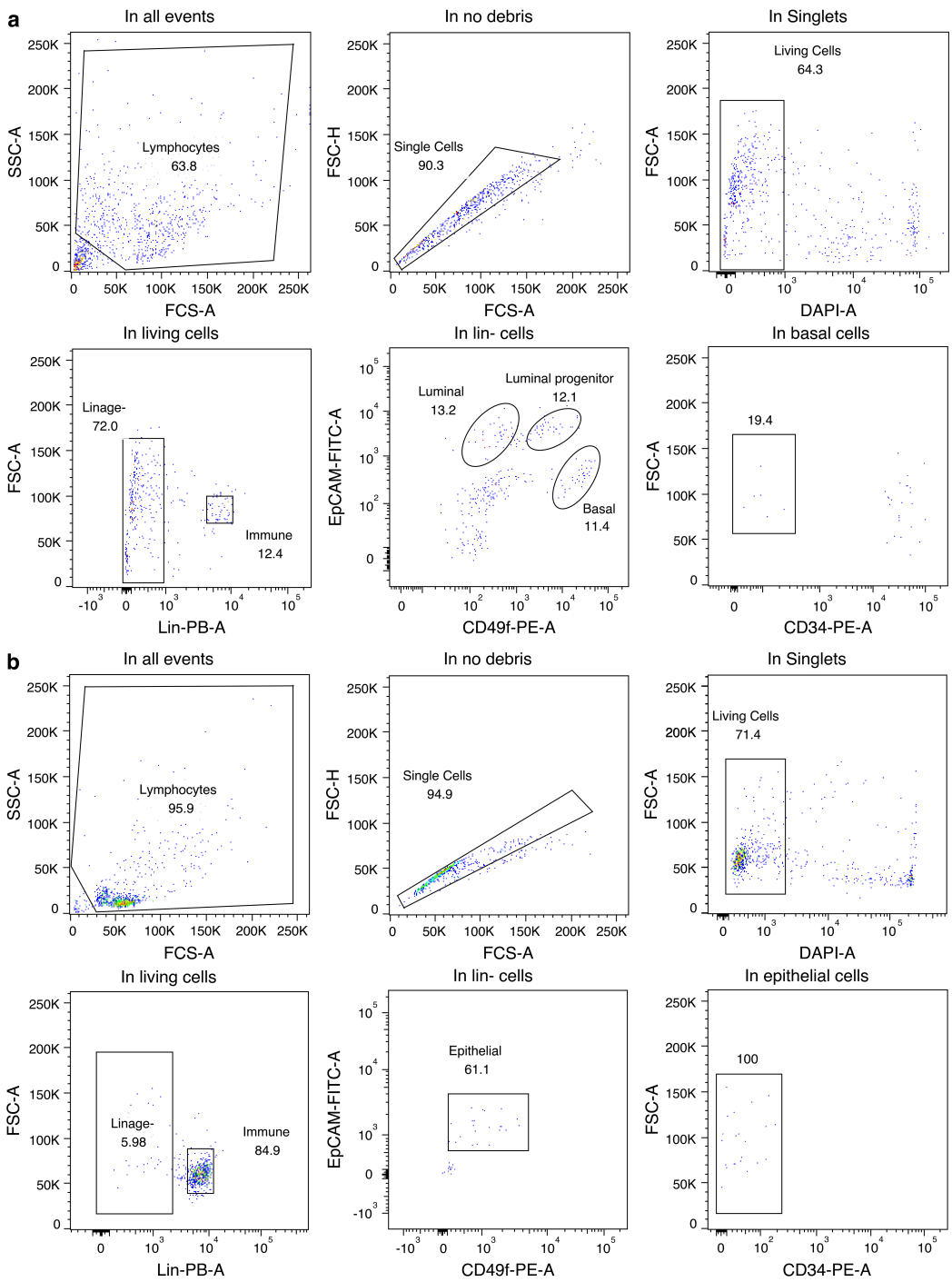
17. Wang, W., Li, W., Wang, C., Fan, W. & Dou, Y. Accurate detection of mosaic mutations at short tandem repeats from bulk sequencing data. *https://github.com/douymLab/BayesMonSTR/tree/main/BayesMonSTR-BulkMonSTR* (2026).

Supplementary Figures

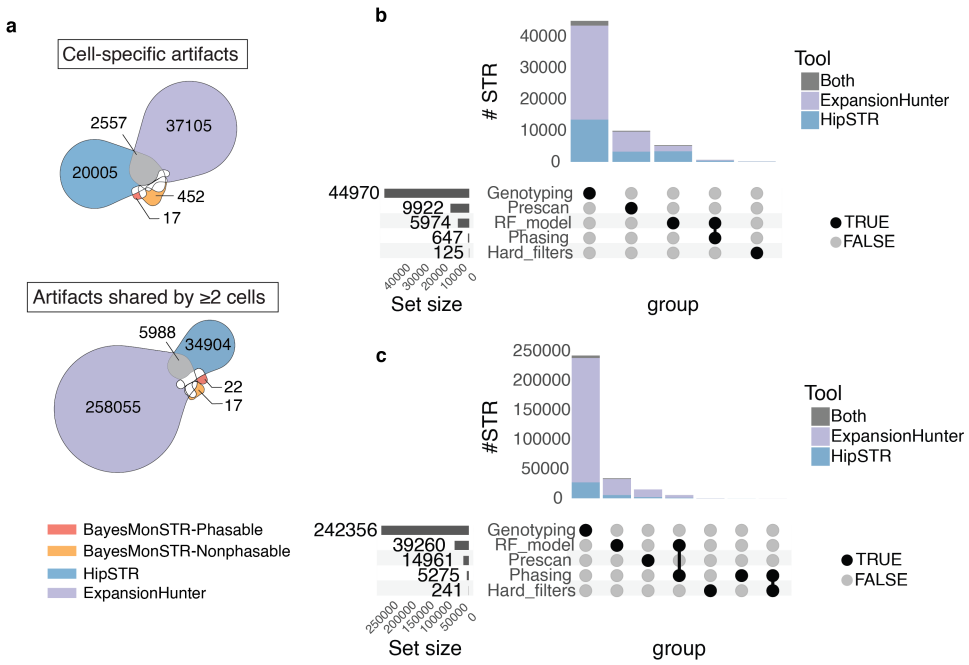


Supplementary Fig. 1 Read segmentation, allele generation and local read alignment based on sequencing reads and information from human reference genome. (a) The repeat-flank region boundary ("anchor" in the Figure) separating the STR-region and flanking regions within sequencing reads often fluctuates due to alignment issues, sequencing errors and other problems. **(b)** A fine-tuned HMM was used for reads segmentation. The three hidden states are "STR", "flanking", and "interruption". Each state exhibits a preference for specific base-pair sequences and emits them with a higher probability. The segmentation boundaries between the STR and two flanking regions were identified by enumerating and selecting the optimal hidden state path. **(c)** STR-specialized local read alignment. Assuming no stutter errors occur, the read alignment likelihood was calculated by multiplying the base agreement probabilities between the read sequences and the hidden allele sequences. In the presence of a stutter error, we assume that DNA slippage occurs only once per read and that the stutter-induced INDEL can occur at any position with equal probability. **(d)** We calculate the probability that each read originates from a specific allele using an STR-specialized local read alignment strategy. **(e-g)** The read

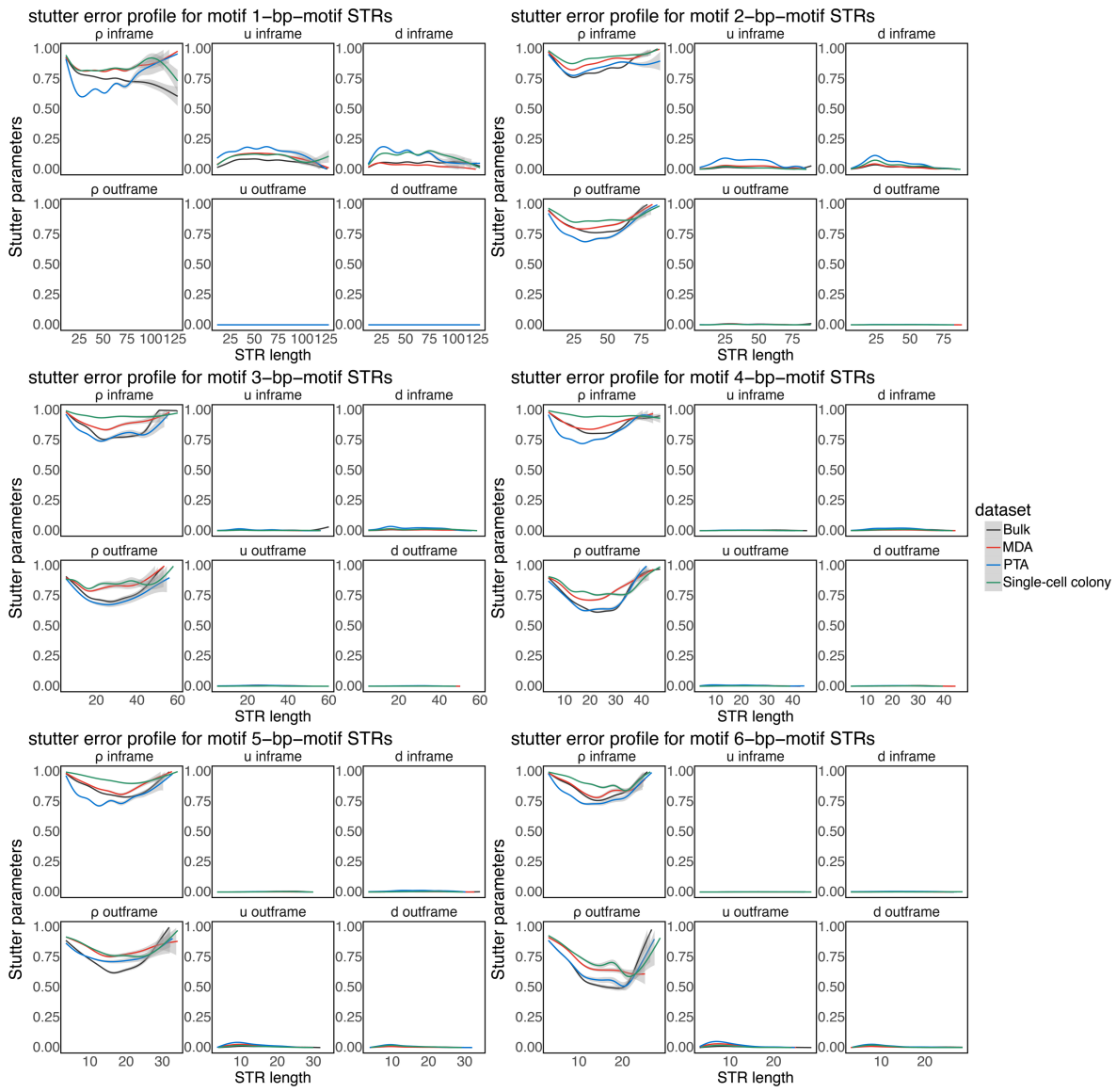
segmentation strategy of BayesMonSTR is robust to sequencing errors and can accurately restore the correct STR length distribution despite indel fluctuations. **(e)** An example demonstrates how the scoring-based Sputnik algorithm, which ignores base quality, leads to segmentation errors. In contrast, BayesMonSTR's HMM-based segmentation strategy incorporates base quality, making it more resilient to sequencing errors. **(f-g)** Compared to BWA alignment, which often struggles with accurately defining STR-flank boundaries, the HMM-based approach of BayesMonSTR effectively tolerates indel fluctuations and restores the correct STR length distribution.



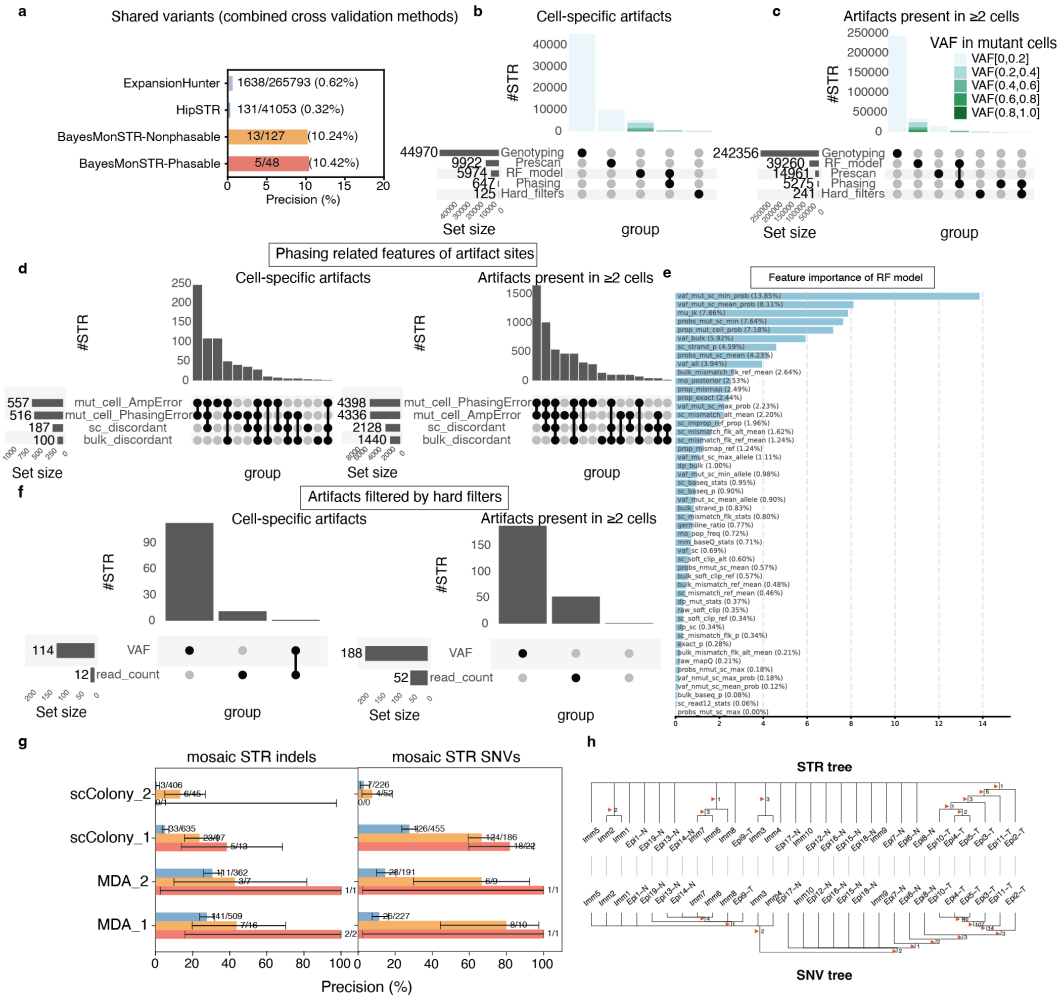
Supplementary Fig. 2 FACS gating strategy for isolating breast-tissue epithelial cells and immune cells. (a) Cell sorting results from paired normal tissue samples. Immune cells were defined as Lin+ cell populations, luminal cells were identified as Lin-EpCAM+CD49f-, luminal progenitor cells were identified as Lin-EpCAM+CD49f+, and basal cells were identified as Lin-EpCAM−CD49f+CD34− cells. **(b)** Cell sorting results from tumor samples. Immune cells were defined as Lin+ cell populations, while epithelial cells were identified as Lin-EpCAM+CD49f+CD34-. The sorting strategy was used to isolate these specific cell subsets for further analysis.



Supplementary Fig. 3 The comprehensive pipeline of BayesMonSTR pipeline effectively excludes artifacts compared with HipSTR and ExpansionHunter. (**a**) Artifact mutations identified by each method for cell-specific (top) and shared (bottom) artifacts from the benchmark dataset we generated. (**b-c**) Artifact mutations from HipSTR and ExpansionHunter were not called by BayesMonSTR through the multiple steps (Genotyping, RF model, Prescan, Phasing, Hard_filters) for cell-specific (**b**) and shared (**c**) artifacts.

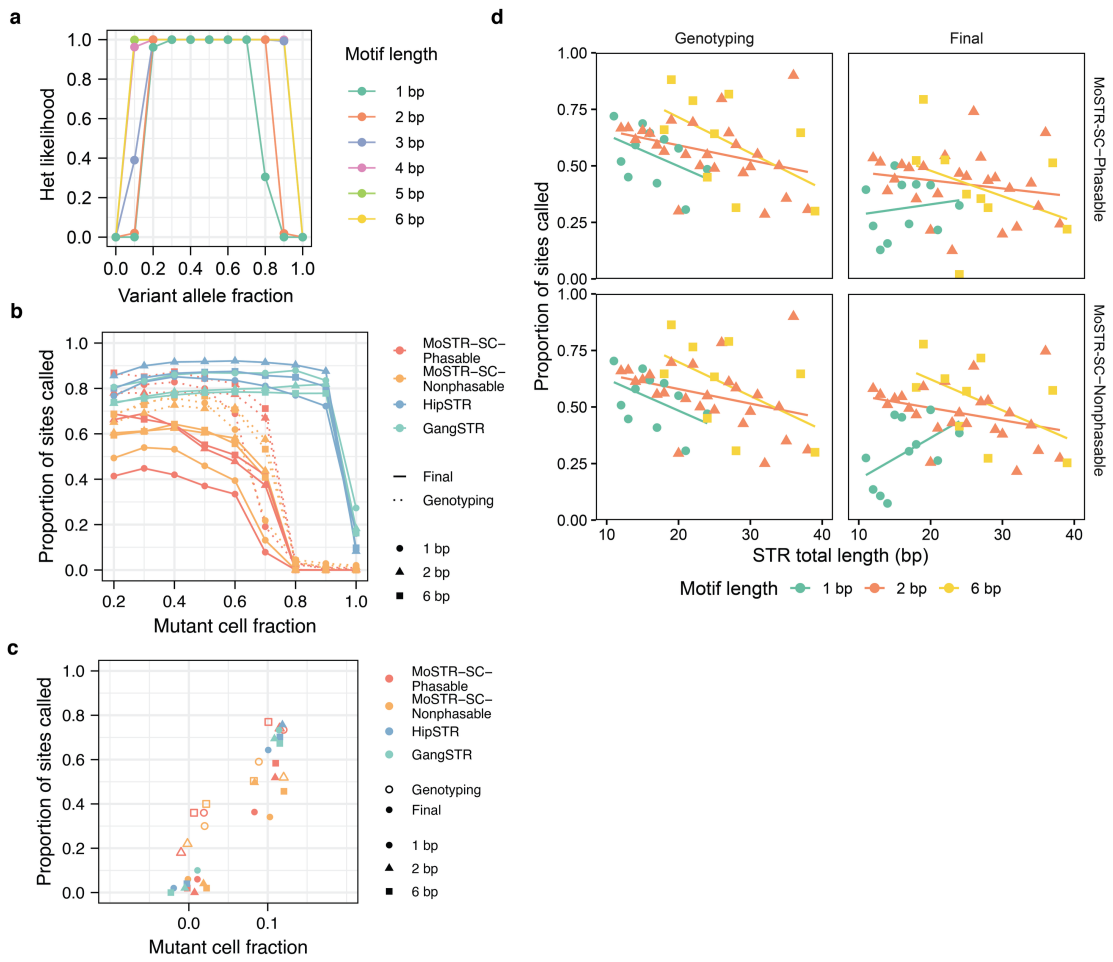


Supplementary Fig. 4 Stutter error estimation across bulk DNA sequencing data, MDA cells, PTA cells and single cell colonies. Stutter-error parameters were derived from a length-based stutter error geometric distribution model, calculated with an EM-algorithm. "Bulk" refers to bulk DNA-seq data, "MDA" refers to single-cell sequencing data amplified with multiple displacement amplification approach, "PTA" refers to single-cell sequencing data amplified with primary template-directed amplification approach, and "SCC" refers to single-cell colony data. STR loci were described by the parameters ρ (rho), u (up), and d (down), which represent the three parameters of the geometric distribution, in-frame and out-frame parameters were estimated separately (Methods). Smoothed trend lines were calculated by the generalized additive model (GAM) with cubic splines, and gray area represents 95% confidence interval.

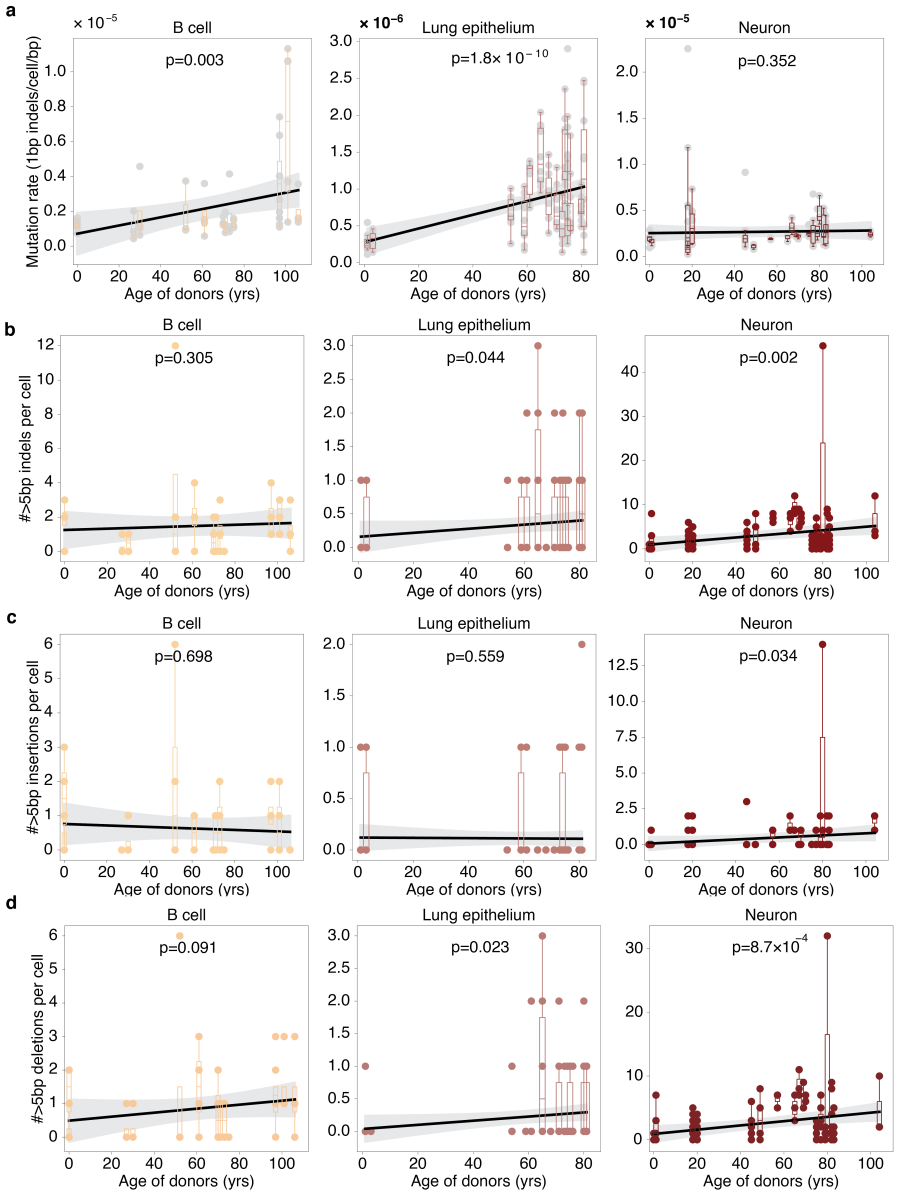


Supplementary Fig. 5 BayesMonSTR outperforms HipSTR and Expansionhunter in detecting mosaic STR mutations from single cells. (a) Validation rate of mutations presents in ≥2 cells across different software tools. The sites were extensively evaluated using both orthogonal data and checked in the phylogeny tree reconstructed with PhyloSOLID (Methods). **(b-c)** Distribution of mutant-cell VAFs for false-positive sites detected by HipSTR and ExpansionHunter. Solid points indicate false positives that were filtered by BayesMonSTR using a specific filtration method. It is evident that BayesMonSTR excluded a majority of these false positives, which had relatively low mutant allele frequencies, during the pre-scan and genotyping stages. Mutations with relatively high mutant-cell VAFs were further excluded by BayesMonSTR using the RF classification model, probabilistic haplotype phasing, and a series of hard filters. **(d)** Expanded analysis of false-positive sites identified by HipSTR and ExpansionHunter excluded by phasing-related features of BayesMonSTR. Features driving the exclusion of false positives by include discordant read pair proportion, amplification error-related features. **(e)** Key features of the random forest model of BayesMonSTR. False positives were effectively filtered by BayesMonSTR by using a combination of informative predictors. Explanations of these predictors are provided in Supplementary Table 1. **(f)** Artifact sites

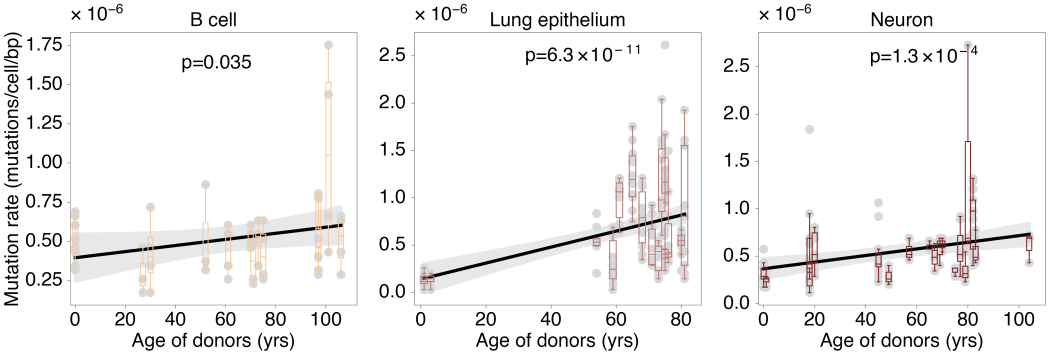
349 excluded by BayesMonSTR using hard filters. **(g)** Performance benchmarking of
350 BayesMonSTR versus HipSTR using additional public datasets. Error bars represent 95%
351 confidence intervals of the validation rates calculated using binomial sampling. MDA_1
352 (UMB1465_21-ctx-1cP2B11_1946-WGS-PFC_19), MDA_2 (UMB1465_21-ctx-
353 1cP2F06_1946-WGS-PFC_20), scColony_1 (10_PRL3-1_001B8_sc16), and scColony_2
354 (10_PRL3-2_002E11_sc17). (**h**) A phylogeny tree reconstructed using shared mutations from
355 our benchmark shows that the mosaic STR tree roughly resembles the mosaic SNV tree. Shared
356 STR variants were identified using only the 16 MDA cells, due to the substantially higher PCR
357 stutter error rate inherent to PTA-amplified cells.



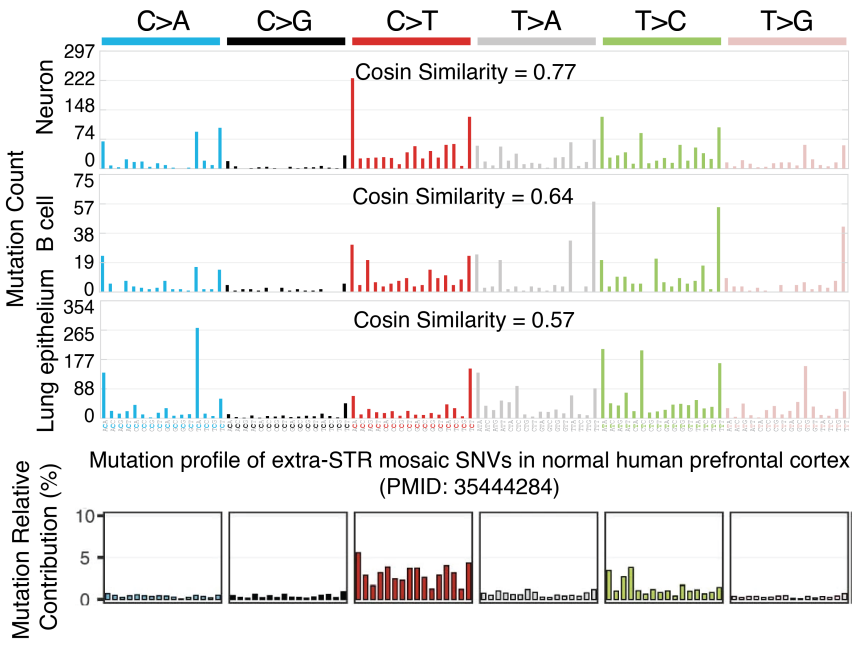
Supplementary Fig. 6 Performance evaluation of BayesMonSTR using simulated data with spike-in mutations. (a) Genotyping likelihoods of germline heterozygous at different variant allele fractions. Reads are assumed with Phred quality score 30 and median stutter error parameters were used for likelihoods calculation, heterozygous likelihoods were normalized along with reference and alternative homozygous likelihoods. **(b-c)** Sensitivity comparison across the two software tools. HipSTR was chosen for the benchmark here, given its superior performance to ExpansionHunter on the real-world data. One-motif mosaic STR insertions within 16-24 bp STR regions were generated using BAMSurgeon. The total number of cells was 10. Dashed lines indicate the raw mutation callings after genotyping of BayesMonSTR, while solid lines indicate final call sets. B illustrates the sensitivity detecting shared mutations and C illustrate cell-specific mutations. **(d)** Sensitivity variation across different STR total lengths. The total cell number was 10 and mutations spiked in were 1-motif insertions.



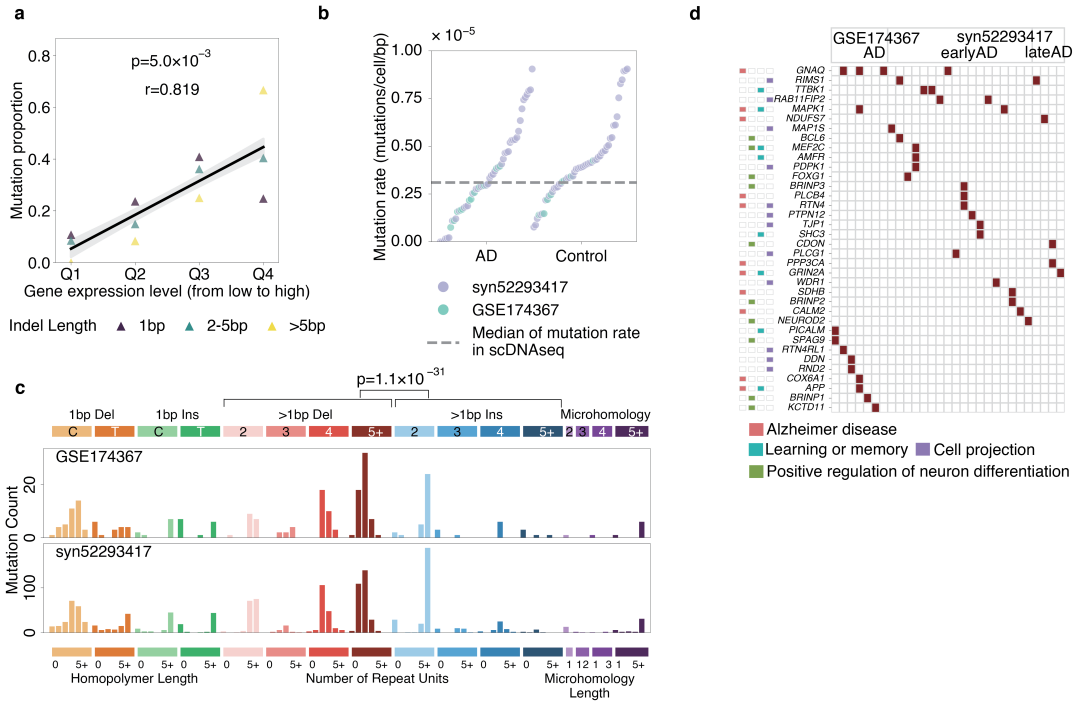
Supplementary Fig. 7 Mosaic STR indels across cell types. (a) 1-bp STR indels accumulate with age in B cells and lung epithelial cells. In contrast, the mutation rate remains relatively stable in non-dividing neurons across donors of different ages. **(b–d)** Relationship between age and the total burden per cell of STR indels >5 bp (b), STR insertions >5 bp (c), and STR deletions >5 bp (d) across three cell types. In all three cell types, the burden of STR deletions >5 bp shows a positive correlation with age. However, only in neurons does the burden of indels, insertions, and deletions consistently increase with age. Linear regression equations were derived using the least squares method. P-values were calculated using a one-tailed t-test to test whether the regression coefficient is significantly greater than zero. The shaded areas around each curve represent the 95% CIs, assuming the differences between predicted values and observed values follow a t-distribution.



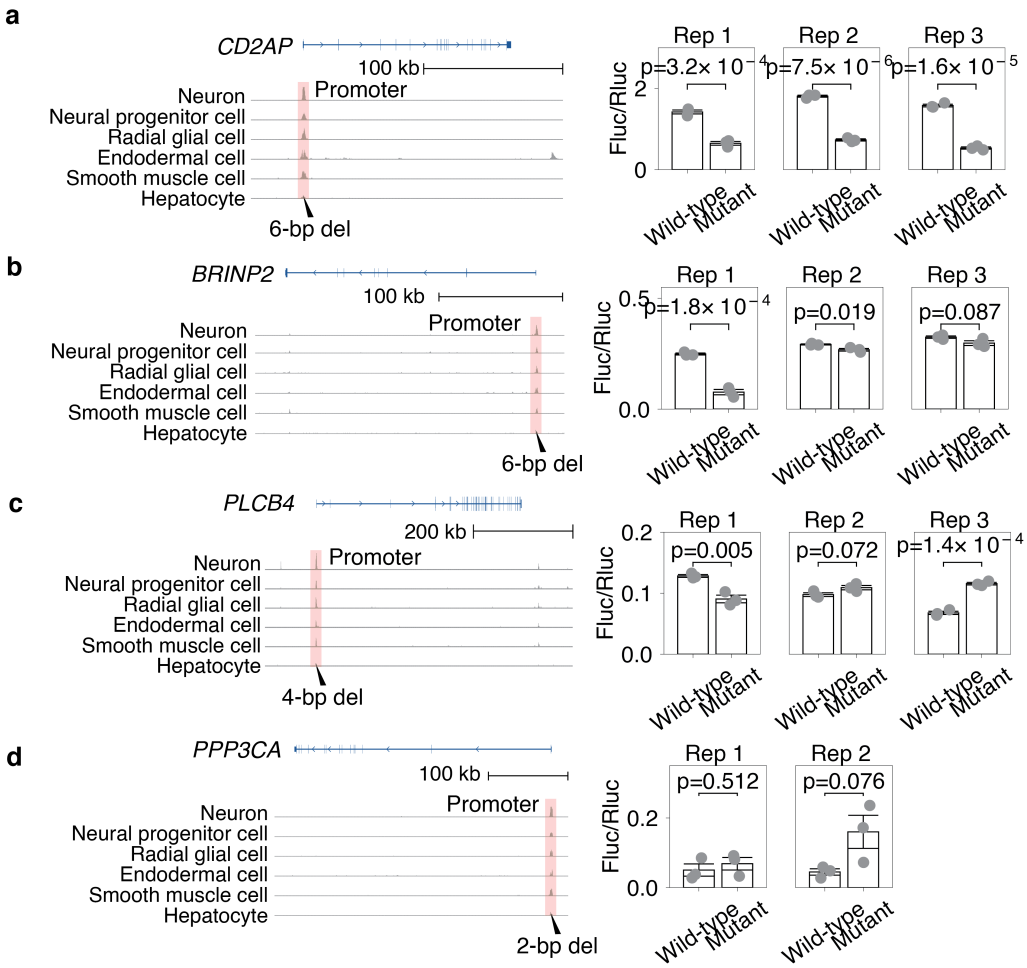
Supplementary Fig. 8 Mosaic STR SNVs across cell types. STR SNVs accumulate with age in all three cell types, as illustrated using the methodology described for Supplementary Figure 7.



Supplementary Fig. 9 Mutational mechanism analysis of mosaic STR mutations. STR-region SNVs largely mirrored the established patterns of extra-STR mosaic SNVs in normal human prefrontal cortex from PMID 35444284. The mutation signature similarity was computed via cosine similarity.



Supplementary Fig. 10 Regulatory mosaic STR indels in neurons. (a) Mosaic STR indels in neurons are enriched in TSSs and active enhancers of highly expressed genes. The number of variants is indicated beside each data point. Q1-Q4 represent expression level quantiles from low to high (Methods), and the brain gene expression data were obtained from ENCODE. Linear regression equations were derived using the least squares method. P-values were calculated using a one-tailed t-test to test whether the regression coefficient is significantly greater than zero. The shaded areas around each curve represent the 95% CIs, assuming the differences between predicted values and observed values follow a t-distribution. **(b)** snATAC-seq samples exhibit mutation rates highly consistent with those observed in scWGS data. **(c)** A significant excess of >1 bp deletions over >1 bp insertions was observed in snATAC-seq data. P-value was calculated using a one-tailed Binomial test. **(d)** Multiple key genes in AD pathways contained multi-hit STR indels in their active regulatory regions across different patients. "earlyAD" and "lateAD" indicate early-stage AD patients and late-stage AD patients.



Supplementary Fig. 11 Luciferase assay results of somatic STR indels in regulatory regions. Normalized levels of H3K4me3 and H3K27ac—key histone modifications linked to promoter and enhancer activities across cell types were obtained from the ENCODE project. Pink regions indicate the promoter or enhancer regions of the gene, with indels highlighted by black triangles. Luciferase assays in HEK293T cells revealed that a 6-bp deletion in the TSS of *CD2AP* and a 6-bp deletion in the TSS of *BRINP2* significantly disrupted gene expression ($P < 0.05$, two-tailed t-test). Furthermore, a 4-bp deletion in the TSS of *PLCB4* and a 2-bp deletion in the TSS of *PPP3CA* enhanced gene expression with marginal significance ($P < 0.1$, two-tailed t-test). Error bars represent standard deviation (3 biological replicates).

Methods

Sequence alignment and data preprocessing

Raw FASTQ reads were aligned to the GRCh37 reference genome (human_g1k_v37_decoy) using BWA-MEM (v0.7.18). Duplicate reads were marked with Picard MarkDuplicates (v2.27), and base quality score recalibration (BQSR) was performed with the Genome Analysis Toolkit (GATK v4.2). Germline variants were joint called from MDA cells and matched bulk sequencing data with GATK HaplotypeCaller and filtered using Variant Quality Score Recalibration (VQSR) following GATK Best Practices. Common germline heterozygous SNPs (gnomAD[1] allele frequency >0.1%) were selected for downstream haplotype phasing with Eagle2 (v2.4.1)[2].

Stutter-error estimation

To reduce amplification- and sequencing-related artifacts introduced by *in vitro* polymerase slippage, BayesMonSTR models STR stutter errors using a geometric-distribution[3]. The stutter error probability $p(r_{m,c}|a=j)$ for m-th read r for cell c in the given allele j was modeled as follows:

$$P(r_{m,c}|a=j) = \begin{cases} 1-u-d, & r_{m,c}=r_a \\ u\rho(1-\rho)^{r_{m,c}-r_a-1}, & r_{m,c}>r_a \\ d\rho(1-\rho)^{r_a-r_{m,c}-1}, & r_{m,c}<r_a \end{cases} \quad (1)$$

Where:

- u denotes the total mutation rate of artifact insertions.
- d denotes the total mutation rate of artifact deletions.
- ρ denotes the success probability of the geometric model, which belongs to the interval (0,1].

Assuming independence across reads, the genotyping likelihood for each diploid cell is computed using equation 2:

$$P(D_c|G_c=(j,k),\theta_{stutter,cell}) = \prod_{m=1}^{n_{reads,c}} \sum_{a\in\{j,k\}} \frac{1}{2} \begin{cases} 1-u-d & r_{m,c}=r_a \\ u\rho(1-\rho)^{r_{m,c}-r_a-1} & r_{m,c}>r_a \\ d\rho(1-\rho)^{r_a-r_{m,c}-1} & r_{m,c}<r_a \end{cases} \quad (2)$$

Where:

- G_c denotes the genotype of a given cell.
- j and k denote the two haplotype alleles.
- D_c denotes the sequencing reads of a given cell c.
- The parameter set θ denotes the stutter model parameters (u, d, ρ), as described in formular (1).

Of note, reads with mapping quality < 20, as well as reads with mean base quality < 20, were excluded before stutter-parameter estimation.

Parameter initialization:

- Given n candidate alleles, the initial population frequency for allele a, f_a, was set to $\frac{1}{n}$, representing equal prior probability across all alleles.
- The initial mosaic allele fraction parameter μ_{msi} was set to 0.001.
- Stutter error parameters u, d and ρ, which define the geometric model, were initialized to 0.01, 0.01, and 0.9, respectively.

E-step:

In the E-step, posterior probabilities of individual germline genotypes were computed using equation 3:

$$\begin{aligned}
&P(G_i^t = (j,k) | G_c^t, D_c, D_{bulk}, \theta^t_{stutter,cell}, \theta^t_{stutter,bulk}) = \\
&P(G_i^t = (j,k)) * P(D_{bulk} | G^t_{bulk} = (j,k), \theta^t_{stutter,bulk}) * \prod_{c=1}^{n_{cells,i}} [P(G_c^t | G_i^t) * P(D_c | G_c^t, \theta^t_{stutter,cell})] = \\
&2 f_j^t f_k^t * \prod_{m=1}^{n_{reads,i,bulk}} \frac{1}{2} \sum_{a \in j,k} P(S_{m,i} = a | \theta^t_{stutter,bulk}) * \prod_{c=1}^{n_{cells,i}} \sum \begin{cases} \frac{1}{2}\mu_{msi} * \sum_{-k \in A_i \setminus \{k\}} P(D_c | G_c^t = (j,-k), \theta^t_{stutter,cell}) \\ (1-\mu_{msi}) * P(D_c | G_c^t = (j,k), \theta^t_{stutter,cell}) \\ \frac{1}{2}\mu_{msi} * \sum_{-j \in A_i \setminus \{j\}} P(D_c | G_c^t = (-j,k), \theta^t_{stutter,cell}) \end{cases}
\end{aligned} \tag{3}$$

Where:

- f_j denotes population allele frequency for allele j.
- G_i denotes the individual germline genotype for individual i, and G_c denotes the cell genotype.
- θ denote the set of stutter error parameters as previously described.
- D_c denotes the sequencing reads of a given cell c.

Here we assume G_i is the same as G_{bulk}, and the prior probabilities of individual genotypes were calculated based on Hardy-Weinberg Theorem. When j equals to k, the prior probability of a homozygous germline genotype is given by f_j^2.

M-step:

We then estimated the error parameters u, d and ρ, and the length-based population allele frequency f at $t+1$-th step for stutter errors in the M-step, and were calculated by equations 4-7, as follows:

$$u^{t+1} = \frac{\sum_{c=1}^{N_{ind}} \sum_{m=1}^{N_{reads,c}} P\left(S_{m,c,i}, r_{m,c,i} | \theta^t_{stutter}\right) I\left(r_{m,c,i} > s_{m,c,i}\right)}{N_{reads}} \tag{4}$$

$$d^{t+1} = \frac{\sum_{c=1}^{N_{ind}} \sum_{m=1}^{N_{reads,c}} P\left(S_{m,c,i}, r_{m,c,i} | \theta^t_{stutter}\right) I\left(r_{m,c,i} < s_{m,c,i}\right)}{N_{reads}} \tag{5}$$

$$\rho^{t+1} = \frac{\sum_{c=1}^{N_{ind}} \sum_{m=1}^{N_{reads,c}} P\left(S_{m,c,i}, r_{m,c,i} | \theta^t_{stutter}\right) I\left(r_{m,c,i} \neq s_{m,c,i}\right)}{\sum_{c=1}^{N_{ind}} \sum_{m=1}^{N_{reads,c}} P\left(S_{m,c,i}, r_{m,c,i} | \theta^t_{stutter}\right) \left|r_{m,c,i} - s_{m,c,i}\right|} \tag{6}$$

$$f_j^{t+1} = \frac{1}{2N_{ind}} \sum_{i=1}^{N_{ind}} \sum_{k=1}^{A_i} \left(P\left(G_i^t = (j,k) \middle| G_{c,i}^t, D_{c,i}, D_{bulk,i}, \theta_{stutter}^t\right) + P\left(G_i^t = (k,j) \middle| G_{c,i}^t, D_{c,i}, D_{bulk,i}, \theta_{stutter}^t\right)\right) \quad (7)$$

Where:

- N_{ind} denotes total number of individuals.
- The indicator function I is used to restrict the computation to reads with insertion, deletion and nonzero stutter for u, d and ρ respectively.
- $p(S, r|\theta)$ is read assignment probability for read $r_{m,c}$ given allele pool A_i (see formula 8).
- G_i denotes the individual germline genotype for individual i, and G_c denotes the cell genotype.
- θ denote the set of stutter error parameters as previously described.
- D_c denotes the sequencing reads of a given cell c.
- N_{reads} denotes the total number of samples with the same sequencing method. (including bulk, WGA, single cell colonies, etc).
- $N_{reads,c}$ denotes the total number of reads for sample c.
- the indicator function I is used to restrict the computation to reads with insertion.
- deletion and nonzero stutter for u, d and ρ respectively.
- θ denoted these stutter parameters.
- f_j denotes population allele frequency for allele j.

$p(S, r|\theta)$ is read assignment probability for read $r_{m,c}$ given allele pool A_i:

$$\begin{aligned}
P(S_{m,c,i}, r_{m,c,i}|\theta_{stutter,cell}^t) &= \sum_{a\in\{A_i\}} P(S_{m,c,i} = a) * P(r_{m,c,i}|S_{m,c,i} = a, \theta_{stutter,cell}^t) = \\
&= \sum_{a\in\{j,k\}} \sum_{j\in\{A_i\},k\in\{A_i\}} P(S_{m,c,i} = a|G_{c,i}^t = (j,k)) * \\
&\quad P(G_{c,i}^t = (j,k)) * P(r_{m,c,i}|S_{m,c,i} = a, \theta_{stutter,cell}^t) = \\
&= \sum_{a\in\{j,k\}} \sum_{j\in\{A_i\},k\in\{A_i\}} \sum_{G_i^t} P(S_{m,c,i} = a|G_{c,i}^t = (j,k)) * \\
&\quad P(G_{c,i}^t = (j,k), G_i^t|D_{c,i}, \theta_{stutter,cell}^t) * P(r_{m,c,i}|S_{m,c,i} = a, \theta_{stutter,cell}^t) = \\
&= \sum_{a\in\{j,k\}} \sum_{j\in\{A_i\},k\in\{A_i\}} \sum_{G_i^t} P(S_{m,c,i} = a|G_{c,i}^t = (j,k)) * \\
&\quad P(G_{c,i}^t = (j,k)|G_i^t) * P(D_{c,i}|G_{c,i}^t = (j,k), \theta_{stutter,cell}^t) * \\
&\quad P(G_i^t) * P(r_{m,c,i}|S_{m,c,i} = a, \theta_{stutter,cell}^t)
\end{aligned} \quad (8)$$

Where:

- a is one of the alleles in the alleles pool $A_i = \{j,k\}$ for individual i.
- other parameters are defined in Equation (3)

Probabilistic read segmentation

Accurate detection of repeat regions within sequencing reads is essential for generating candidate STR alleles. Traditional segmentation methods, which rely on alignment coordinates and repeat patterns, often suffer from inaccuracies due to mapping biases

and interruptions. These issues lead to fluctuations in the 'anchor points' used to distinguish repeat and flanking regions, causing the anchor points to deviate from their expected coordinate across different reads (Extended Data Fig. 1a). To address these issues, we utilized a Hidden Markov Model (HMM) approach to identify repeat regions in sequencing reads.

Notably, our HMM model integrates information from both sequencing reads (including sequencing data and quality scores) and the reference genome (repeat-region and flanking region sequences). It also incorporates prior knowledge of in-frame and out-of-frame interruptions derived from the human reference genome. This design enables robustness to sequence interruptions, sequencing errors, and mutations within STR regions, while integrating base quality scores directly into the segmentation process (Extended Data Fig. 1).

Specifically, the HMMs were designed to model the STR sequences using three hidden blocks (Extended Data Fig. 1b):

A. STR block: Represents perfect repeat motifs, capturing the repetitive structure at each STR locus. For example, the motif CAG is modeled with three hidden states—S_C, S_A, and S_G —within this block.
B. Interruption block: Models mismatches and indels within repeat regions. Similar to profile HMM[4], this block accounts for variations inside the repeat tract.
C. Flanking block: Represents sequences flanking the STR. In addition to a general flanking state with equal emission probabilities for A, C, G and T, we introduce a locus-specific 3-bp known flanking states extracted from the human reference genome. These states incorporate sequence context to improve segmentation accuracy.

To initialize the transition probabilities, the initial transition probability from the STR block to the flanking block was defined as 1/L, where L corresponds to the expected segmentation distance before entering the flanking region. For homopolymers, L was set to 5 bp, while for non-homopolymers, L was set to twice the motif length. Furthermore, the initial transition probability from the STR block to the interruption block is estimated through an extensive grid search using a reference panel[3].

The emission probability matrix integrates both the expected STR motif sequence and base quality scores to improve robustness against sequencing errors. Each hidden state corresponds to an expected nucleotide that receives the highest emission probability. In the flanking block, known 3-bp flanking sequences from the reference genome are assigned predefined emission probabilities, while nucleotides outside this region have equal probabilities across A, C, G, and T. For a given hidden state, if the expected base is X_t, the observed base is Y_t, and the sequencing error rate is e, the emission probability for this base is calculated as follows:

$$P(Y_t \mid X_t) = \begin{cases} 1-e, & Y_t = X_t \\ \frac{e}{3}, & Y_t \neq X_t \end{cases} \tag{9}$$

Where:

- Y_t indicates the observed base in sequencing reads.
- X_t indicates the expected base given a specific hidden state.
- e indicates sequencing error rate.

Finally, the boundaries between STR and flanking regions are determined by finding the maximum-likelihood hidden state path using the Viterbi algorithm. Based on this optimal state path, sequencing reads are segmented into STR and flanking segments. In cases of large deletions where the mutant allele length falls below the predefined step length—leading to potential segmentation failure—a BWA-based coordinate mapping approach is used as a backup. Similarly, due to the high sequence complexity at imperfect STR loci with 4-bp, 5-bp, or 6-bp motifs containing interruptions, segmentation performance may be compromised; for such loci, the BWA-based coordinate method is also applied.

STR-specific read alignment

STR sequences from the segmented reads are used to generate candidate alleles, each comprising the core STR region and 5 bp of flanking sequence on both sides. Segmented reads are subsequently aligned to the candidate alleles using a STR-specific alignment method[3] to calculate the alignment likelihood. This method accounts for both potential PCR stutter errors and sequencing errors. Specifically, three possible scenarios are considered for the sequencing reads: no stutter error, stutter deletion, and stutter insertion. The calculation of alignment likelihood varies slightly for each scenario, as illustrated in Extended Data Fig. 1c.

Given a haplotype X of length L, if read Y is assumed to be free of stutter errors, the likelihood of the observed alignment is determined by the agreement between each base i in the read and its corresponding base in the haplotype:

$$P(X,Y) = (1-u-d) * \prod_{i=1}^{L} Q(X_i, Y_i) \tag{10}$$

Where:

- X denotes the given STR haplotype.
- Y denotes the STR haplotype observed in the read.
- L denotes the STR length of X haplotype.
- u denotes stutter insertion rate from the estimated stutter model.
- d denotes stutter deletion rate from the estimated stutter model.
- $Q(X_i, Y_i)$ represents the agreement score between the i-th base of haplotype X and the i-th base of haplotype Y,as defined in Equation (9).

Otherwise, if read Y contains a stutter deletion of haplotype X—with a deletion size of ΔL—we assume that the deletion could occur at any position k within the haplotype, where k ranges from 1 to total_haplotype_length $-\Delta L + 1$. The read alignment likelihood is then determined jointly by the probability of stutter deletion and the

base-wise agreement between the read and the haplotype. The overall likelihood is obtained by averaging over all possible deletion positions:

$$P(X,Y) = d\rho(1-\rho)^{\frac{\Delta L}{L_{RU}}-1} * \frac{1}{L-\Delta L+1} \sum_{k=1}^{L-\Delta L+1} \prod_{i=1}^{k-1} Q(X_i, Y_i) \prod_{i=k+\Delta L}^{L} Q(X_i, Y_{i-\Delta L}) \quad (11)$$

Where:

- ΔL denotes deletion size in read Y.
- L_{RU} denotes the motif length (in bp).
- ρ denotes the success probability of the geometric model, which belongs to the interval (0,1].
- k denotes the deletion position in the haplotype X.
- Other parameters are defined in Equation (10).

In the third scenario, where the read Y results from a stutter insertion in haplotype X (insertion size ΔL), we assume the inserted sequence arises from a local duplication of adjacent sequence. Consequently, the duplication can originate from either the left or right flank of the insertion site. Due to boundary constraints—insertions within the first ΔL positions can only be duplicated from the right flank, and those within the last ΔL positions only from the left—only one duplication direction is considered for such edge cases. The alignment likelihood is determined jointly by the stutter insertion rate and the base-wise agreement between the read and the haplotype. As before, the overall likelihood is obtained by averaging over all possible insertion positions and duplication directions:

$$P(X,Y) = \left(u*\rho*(1-\rho)^{\frac{\Delta L}{L_{RU}}-1}*\frac{1}{L+1}\right)*\sum \begin{cases} \sum_{k=0}^{\Delta L-1} \left(\prod_{i=1}^{k+\Delta L} Q(X_i, Y_i)\right) * \left(\prod_{i=k+1}^{L} Q(X_i, Y_{i+\Delta L})\right) \\ \sum_{k=\Delta L}^{L-\Delta L} \left(\prod_{i=1}^{k} Q(X_i, Y_i)\right) * \frac{1}{2}\left(\prod_{i=k+1}^{k+\Delta L} Q(X_i, Y_i) + \prod_{i=k+1}^{k+\Delta L} Q(X_{i-\Delta L}, Y_i)\right) * \left(\prod_{i=k+1}^{L} Q(X_i, Y_{i+\Delta L})\right) \\ \sum_{k=L-\Delta L+1}^{L} \left(\prod_{i=1}^{k} Q(X_i, Y_i)\right) * \left(\prod_{i=k+1}^{k+\Delta L} Q(X_{i-\Delta L}, Y_i)\right) * \left(\prod_{i=k+1}^{L} Q(X_i, Y_{i-\Delta L})\right) \end{cases} \quad (12)$$

Where:

- ΔL denotes the insertion size in read Y.
- L_{RU} denotes the motif length.
- ρ denotes the success probability of the geometric model, which belongs to the interval (0,1].
- k denotes the insertion position in the haplotype X.
- Other parameters are defined in Equation (10).

We use alignment likelihoods to estimate the probability that each read originates from a given haplotype, thereby determining the most likely haplotype of origin for each read (Extended Data Fig. 1d). The mutation type (insertion, deletion, or mismatch) was further determined by pairwise alignment of the inferred mutant haplotype to the germline haplotype using pairwise aligner[5] (v1.81). This refinement alignment incorporated 4-bp flanking sequences on each side.

Estimating global whole-genome amplification bias

To estimate whole-genome allelic balance for each single cell, we modeled alternative allele counts using a beta-binomial distribution. Germline heterozygous SNPs (hSNPs) were defined as sites with matched-bulk VAF from 0.4 to 0.6 and present in dbSNP. The two shape parameters of the Beta distribution, α and β were estimated as follows:

$$\alpha = \frac{dp * m_1 - m_2}{dp * (\frac{m_2}{m_1} - m_1 - 1) + m_1} \tag{13}$$

$$\beta = \frac{(dp - m_1) * (dp - \frac{m_2}{m_1})}{dp * (\frac{m_2}{m_1} - m_1 - 1) + m_1} \tag{14}$$

where:

- $m_1 = mean(n_{alt})$.
- $m_2 = \frac{\sum_{m=1}^{n_{site}} n_{alt,m}^2}{n_{site}}$.
- dp is defined as the median depth of hSNP loci.

Notably, the hSNPs collected for calculating allele imbalance parameters were all of the same median depth, such as 20x. n_{alt} refers to the alternative read count, and n_{site} refers to the total number of hSNP loci.

Estimating local amplification bias using a piecewise gaussian process regression

In whole-genome amplification, uneven amplification across DNA fragments creates locally correlated allelic bias. We addressed this by segmenting reads into fragments using an HMM and then applying piecewise Gaussian process regression to quantify bias within each segment.

To model the correlation of allelic balance, we implemented a piecewise Gaussian process regression framework. First, reads within individual cells were partitioned into independent segments using a hidden Markov model (HMM). Within each segment, the haploid allelic fraction (AF_h) values of the same haplotype h, derived from phased hSNPs (phased with Eagle2 as described previously), along with their genomic coordinates, served as inputs for the regression. The Gaussian process is formally defined as:

$$B_h = f(x) \sim GP(\mu = 0, K(x, x'))$$
$$K(x, x') = \sigma_{frag}^2 \exp(-\frac{(x - x')^2}{2l_{frag}^2}) \tag{15}$$

and the allele fraction is represented using a logistic transform:

$$AF_h = \frac{1}{1 + e^{-B_h}}, \quad B_h \in (-\infty, +\infty) \tag{16}$$

Where:

- allele imbalance B_h given hSNPs is modelled by Gaussian process GP with zero mean function.

- the radial basis function $K(x, x')$ with parameter σ and l accounting for fragments attribution.

To prevent discontinuities in the regression line caused by allele dropout, the haploid allelic fraction (AF_h) was constrained to the interval (10^{-7}, $1 - 10^{-7}$). This model was subsequently applied to predict AF values at each candidate STR loci. Genomic regions containing phased SNPs were excluded if their correlation curves showed sharp fluctuations, specifically those with a mean squared error (MSE) exceeding 0.5, to filter out intervals with potential phasing errors.

Calculation of mosaic genotyping likelihood with bulk sequencing data

We made no specific assumption about the allele frequencies of mosaic mutations; instead, we adopted a uniform prior distribution over the interval [0, 1]. Based on this, the mosaic genotyping likelihood given bulk sequencing data can be calculated with the following formula:

$$P(D|G_i = mosaic) = \int_0^1 (\theta)^r (1-\theta)^{depth-r} d\theta = \binom{depth}{r} \beta(r+1, depth-r+1) \qquad (17)$$

Where:

- r indicates alternative read count in the bulk sequencing data.
- $depth$ indicates total read depth in the bulk seq data.
- θ indicates the prior distribution of variant allele frequency of mosaic mutations.

Joint Genotyping

Based on population allele frequencies, sequencing reads and stutter error profiles, BayesMonSTR calculated genotyping posteriors of candidate mosaic STR mutations with a haplotype-based EM-algorithm.

Parameter initialization:

Population frequencies f_a were set to be with equal probability. Haplotype phasing probability $P(j - h_1, k - h_2|G_i)$ was set to 0.5. μ_{msi} was set to 0.001. Stutter error parameters u, d and ρ for each site were calculated with previous procedures.

E-step:

BayesMonSTR employed a hierarchical Bayesian network to incorporate population, bulk and single cell-level data. We utilized an EM algorithm to calculate maximum likelihood of latent parameters.

Similar to the length-based EM procedure, the E-step for estimating individual germline genotypes at sequence-level resolution was performed as follows:

$$\begin{aligned} &P(G_i^t = (j,k)|D_c, D_{bulk}) = P(G_i^t = (j,k)) * P(D_{bulk}|G_{bulk}^t = (j,k)) * \prod_{c=1}^{n_{cells,i}} \sum_{G_c^t} [P(G_c^t|G_i^t = (j,k)) * P(D_c|G_c^t)] = \\ &2 f_j^t f_k^t * \prod_{m=1}^{n_{reads,bulk}} \sum_{a \in \{j,k\}} \frac{1}{2} P(r_{m,i}, Q_{m,i}|S_{m,i} = a) * \prod_{c=1}^{n_{cells,i}} \sum \begin{cases} \sum_{\neg k \in A_i \setminus \{k\}} \mu_{i,k\to\neg k}^t * (1 - \sum_{\neg j \in A_i \setminus \{j\}} \mu_{i,j\to\neg j}^t) * P(D_c|G_c^t = (j,\neg k)) \\ (1 - \sum_{\neg k \in A_i \setminus \{k\}} \mu_{i,k\to\neg k}^t) * (1 - \sum_{\neg j \in A_i \setminus \{j\}} \mu_{i,j\to\neg j}^t) * P(D_c|G_c^t = (j,k)) \\ \sum_{\neg j \in A_i \setminus \{j\}} \mu_{i,j\to\neg j}^t * (1 - \sum_{\neg k \in A_i \setminus \{k\}} \mu_{i,k\to\neg k}^t) * P(D_c|G_c^t = (\neg j,k)) \end{cases} \end{aligned} \qquad (18)$$

If the germline genotype is homozygous, the prior probability of G_i is f_j^2 instead of $2f_jf_k$.

Equation 2 and 19 were used for unphased and phased genotype likelihood calculation, respectively.

$$
\begin{aligned}
P(D_c|G_c^t=(j,k)) &= \prod_{m=1}^{n_{reads,c}} \sum \begin{cases} P(O_{m,c}|S_{m,c}=j, S_{m,c}=h_1) * P_{j-h_1}^t \\ P(O_{m,c}|S_{m,c}=j, S_{m,c}=h_2) * P_{j-h_2}^t \\ P(O_{m,c}|S_{m,c}=k, S_{m,c}=h_1) * P_{j-h_2}^t \\ P(O_{m,c}|S_{m,c}=k, S_{m,c}=h_2) * P_{j-h_1}^t \end{cases} \\
&= \prod_{m=1}^{n_{reads,c}} \sum \begin{cases} AF_{c,j-h_1} * P(r_{m,c}|a=j) * P_{j-h_1}^t \\ AF_{c,j-h_2} * P(r_{m,c}|a=j) * P_{j-h_2}^t \\ AF_{c,k-h_1} * P(r_{m,c}|a=k) * P_{j-h_2}^t \\ AF_{c,k-h_2} * P(r_{m,c}|a=k) * P_{j-h_1}^t \end{cases}
\end{aligned}
\tag{19}
$$

Where:

- $AF_{c,j-h_1}$ denotes the predicted allele fraction based on STR allele j at the locus for cell c.
- P_{j-h_1} denotes the phasing probability that STR allele j is linked with germline hSNP h_1.
- $P(r_{m,c}|a=j)$ denotes the read likelihoods given allele j based on the stutter error model.

M-step:

We calculated the phasing probabilities that quantify the phasing relationships between STR loci and hSNPs, which was initialized with 0.5. If STR allele j linked with germline SNP h_1 and allele k links with germline SNP h_2, the phasing probability given individual i read information D_i was calculated with the following equation:

$$
\begin{aligned}
P^{t+1}(j-h_1,k-h_2|D_i) &= \sum_{G_i^t} P^t(j-h_1,k-h_2|D_c,D_{bulk},G_i^t) * P(G_i^t) \\
&= \sum_{G_i^t} \frac{P(j-h_1,k-h_2|G_i^t) * P(G_i^t) * P(D_{bulk}|j-h_1,k-h_2,G_i^t) * P(D_c|j-h_1,k-h_2,G_i^t)}{P(D_c,D_{bulk}|G_i^t)} \\
&= \sum_{G_i^t} \frac{P(j-h_1,k-h_2|G_i^t) * P(G_i^t) * P(D_{bulk}|j-h_1,k-h_2,G_i^t) * \prod_{c=1}^{n}[P(D_c|j-h_1,k-h_2,G_c^t=(j,k)) * P(G_c^t=(j,}{P(D_{bulk}|G_i^t) * \prod_{c=1}^{n}[P(G_c^t=(j,k)|G_i^t) * P(D_c|G_c^t=(j,k))]}
\end{aligned}
\tag{20}
$$

Where:

- $P(G_i^t)$ denotes the population haplotype frequency prior.
- $p(D \mid j-h_1, k-h_2)$ is the genotype likelihood for individual i given allele j linked with SNP h_1 and allele k links with SNP h_2, and with $P(j-h_1,k-h_2|G_i^t)$ set to 0.5.
- $P(G_c^t = (j,k)|G_i^t)$ denotes the mosaic mutation probability that the cell genotype is (j,k) given G_i.

Similarly, the probability that STR allele j is linked with germline SNP h_2 follows the same approach. Of note, we only used reads covering both the candidate locus and the nearby hSNP to calculate the phasing probability.

The mosaic allele fraction for mutant allele $\neg k$ was calculated using the following function:

$$\mu_{i,k\to\neg k}^{t+1} = \frac{\sum_{j\in A_i\setminus\{k\}} P(G_i^t=(j,k))\sum_c P(G_i^t=(j,\neg k)|G_i^t=(j,k)) + P(G_i^t=(k,k))\sum_c P(G_i^t=(k,\neg k)|G_i^t=(k,k))}{\sum_{j\in A_i\setminus\{k\}} P(G_i^t=(j,k))\sum_c 1 + P(G_i^t=(k,k))\sum_c 2} \tag{21}$$

Where:

- $(G_i^t = (j,k))$ denotes the individual germline probability in E-step.
- k denotes the germline allele.

The sequence-based STR allele population frequency was calculated as follows:

$$f_j^{t+1} = \frac{1}{2N_{ind}} \sum_{i=1}^{N_{ind}} \sum_{k=1}^{A_i} \left(P\left(G_i^t=(j,k)\middle|G_{c,i}^t, D_{c,i}, D_{bulk,i}, \theta_{stutter}^t\right) + P\left(G_i^t=(k,j)\middle|G_{c,i}^t, D_{c,i}, D_{bulk,i}, \theta_{stutter}^t\right)\right) \tag{22}$$

Where:

- All parameters are defined in Equation (18).

Estimating somatic genotyping posteriors for each site across cells:

BayesMonSTR employed a dynamic programming approach to compute the posterior probability that a candidate locus contains at least one mutant cell. The posterior probability that a mosaic mutation occurs ($k > \neg k$) is calculated with the following formula:

$$P_{i,k\to -k} = 1 - P\left(L_{i,k\to -k} = 0|\bar{D}_c\right) \tag{23}$$

Where:

- L_i denotes the mutant allele count for individual i.

The calculation follows with Bayes' rule:

$$\begin{aligned}
P\left(L|\bar{D}_c\right) &= \frac{P(L) * P\left(\bar{D}_c|L\right)}{P\left(\bar{D}_c\right)} \\
&= \frac{P(L) * \sum_{\vec{G}_c} P\left(\bar{D}_c|\vec{G}_c, L\right) P\left(\vec{G}_c|L\right)}{\sum_{L'=0}^{N} P(L') \sum_{\vec{G}'_c} P\left(\bar{D}_c|\vec{G}'_c, L'\right) P\left(\vec{G}_c|L'\right)} \\
&= \frac{P(L) * \sum_{\vec{G}_c} P\left(\bar{D}_c|\vec{G}_c\right) P\left(\vec{G}_c|L\right)}{\sum_{L'=0}^{N} P(L') \sum_{\vec{G}'_c} P\left(\bar{D}_c|\vec{G}_c\right) P\left(\vec{G}_c|L'\right)} \\
&= \frac{P(L) * \sum_{\vec{G}_c} P\left(\bar{D}_c|\vec{G}_c\right) \delta_{L,S(\vec{G}_c)} \frac{1}{\binom{N}{L}}}{\sum_{L'=0}^{N} P(L') \sum_{\vec{G}'_c} P\left(\bar{D}_c|\vec{G}_c\right) * \delta_{L',S(\vec{G}'_c)}} \\
&= \frac{\frac{1}{\binom{N}{L}} * P(L) * \sum_{\vec{G}_c} \delta_{L,S(\vec{G}_c)} \prod_c P(D_c|G_c)}{\frac{1}{\binom{N}{L}} * \sum_{L'=0}^{N} P(L') \sum_{\vec{G}'_c} \delta_{L',S(\vec{G}'_c)} \prod_c P(D_c|G'_c)}
\end{aligned} \tag{24}$$

The Kronecker delta function, denoted as $\delta_{L,S}$ is a mathematical function that evaluates to 1 when its indices are equal, and evaluates to 0 when its indices differ.

$$P(L_{i,k\to\neg k}) = \binom{N_c}{L_{i,k\to\neg k}} \mu_{i,k\to\neg k}^{L_{i,k\to\neg k}} (1-\mu_{i,k\to\neg k})^{N_c - L_{i,k\to\neg k}} \tag{25}$$

The recursive formula of dynamic programming is as follows:

$$M_{l,j} = \begin{cases} \sum_{\vec{G}_c} P\left(\bar{D}_c|\vec{G}_c\right) \delta_{l,S(\vec{G}_c)} \prod_{c=1}^{j} P(D_c|G_c) & 0 \le l \le j \\ 0 & l > j \end{cases} \tag{26}$$

$$M_{l,j} = M_{l,j-1} * P(D_j|G_j = (j,k)) + M_{l-1,j-1} * P(D_j|G_j = (j,-k)) \tag{27}$$

Where:

- $M_{l,j}$ denotes probability for mutant allele count l and allele j in the DP matrix.

Haplotype phasing related artifacts

Haplotype phasing is widely adopted as a methodology to increase confidence of mosaic mutation calling[6]. The probabilistic haplotype phasing approach we designed is based on two key assumptions: 1) Given a relatively low mutation rate, it is highly unlikely that somatic mutations would occur independently on both haplotypes at the same position; 2) non-cancer cells possess diploid genomes.

Assuming STR allele j is linked with hSNP h_1, and STR allele k is linked with hSNP h_2. Discordant rate of reads for bulk sample d_{bulk} and single cell sequencing data d_{sc} were defined as

$$d_{bulk} = \frac{n_{k-h_1}}{n_{j-h_1} + n_{k-h_1}} \tag{28}$$

$$d_{sc} = \frac{n_{k-h_1}}{n_{j-h_1} + n_{k-h_1}} \tag{29}$$

Amplification error d_{amp} rate and discordant rate d_k for mutant cells sample were defined as:

$$d_{amp} = \frac{n_{j-h_2}}{n_{j-h_2} + n_{k-h_2}} \tag{30}$$

$$d_k = \frac{n_{k-h_1}}{n_{k-h_1} + n_{k-h_2}} \tag{31}$$

Where:

- n_{j-h_1} is the number of reads that STR allele j is linked with hSNP h_1.
- n_{k-h_1} is the number of reads that STR allele k is linked with hSNP h_1.
- n_{j-h_2} is the number of reads that STR allele j is linked with hSNP h_2.
- n_{k-h_2} is the number of reads that STR allele k is linked with hSNP h_2.

Based on empirical assessment, candidate mutations are considered high-confidence somatic events when all four of their phasing error rates (defined in Equations 28–31) are below 0.1.

Genome-wide variant prediction using Random Forest (RF) classification model

We trained two Random Forest classifiers to predict mosaic variants using a rigorously curated training set with phasing information (Supplementary Table 1). The first model incorporated phasing-related features (Supplementary Table 1), while the second model was designed for genome-wide prediction, regardless of phasability (Supplementry Table 1). A leave-one-donor-out strategy was applied in model training and testing.

713 The model was specifically trained to predict refined genotypes, categorizing sites into
714 five distinct classifications: amplification error (amp error), sequencing error (seq error),
715 repeat, heterozygous (het) and mosaic. To ensure an unbiased performance assessment,
716 we employed a leave-one-out cross-validation strategy for model training and testing.
717 To identify mosaic loci, we applied a dual-mode filtering strategy based on phasing
718 feasibility. Phase-informed mode: Variants initially predicted by the phase Random
719 Forest model underwent a series of phasing-based hard filters to yield a high-confidence
720 set of phasing-corrected mosaic loci. These filters included thresholds for bulk, single-
721 cell, and mutant-cell discordance rates, alongside an amplification error rate (all < 0.1,
722 as defined in Equations 28–31), supplemented by requirements for mutant cell VAF $\geq$
723 0.1 and a mutant read count ≥ 2. Non-phaseable mode: For variants lacking reliable
724 phase information, two sequential Random Forest models were applied to mitigate
725 amplification errors. To derive a stringent set of true mosaic sites, we then applied an
726 additional empirical filter requiring mutant cell VAF > 0.25 and mutant read count ≥ 3.

727 **Examination of artifact sites identified by different software tools**

728 To investigate the superior performance of BayesMonSTR over HipSTR and
729 ExpansionHunter in detecting mosaic STR mutations from single-cell data, we
730 systematically analyzed the false-positive calls from the latter two tools. This analysis
731 revealed how BayesMonSTR's multi-step filtering pipeline eliminates such artifacts.
732 The key filtering steps are defined as follows: 1. Prescan: Candidate loci lacking
733 alternative alleles or residing in low-depth regions (mean depth ≤ 1 across cells) were
734 removed. 2. Genotyping: Candidate mutations with a mosaic genotyping posterior
735 probability below 0.9 were excluded. 3. Phasing: Candidate mutations were filtered
736 based on the phasing metrics defined in Equations 28–31. Specifically, we excluded
737 any mutation with a bulk discordant rate, single-cell discordant rate, amplification error
738 rate, or mutant-cell discordant rate exceeding 0.1. 4. RF model: Candidate mutations
739 were classified as mosaic if the highest predicted probability was for the mosaic class.
740 The SHAP[7] (SHapley Additive exPlanations, SHAP, v0.46.0) method was then used to
741 identify the key features driving the RF classification for these mutations (Extended
742 Data Fig. 4e).

743 **Simulation of spike-in mosaic STR mutations in single cell sequencing data**

744 To simulate mosaic STR mutations in single cell sequencing data, STR indels and
745 mismatches were randomly introduced at target loci using BAMSurgeon[8] (v1.4.1).
746 Diploid single-cell data were then sampled via beta-binomial sampling to emulate
747 allele-specific amplification imbalance. Finally, aligned reads from 10 single-cell
748 libraries (male donors) were down-sampled and merged to generate synthetic diploid
749 datasets for validation.

750 **Breast cancer sample collection**

751 A fresh breast cancer tissue sample and a matched normal breast tissue were obtained
752 from a female breast cancer patient at The Second Affiliated Hospital Zhejiang
753 University School of Medicine. The study was approved by the Institutional Review

754 Board of Westlake University (IRB: 20230213DYM001). Informed consent was
755 obtained from the patient in accordance with the IRB-approved protocol, and consent
756 for data publication was also provided. At the time of collection, pathologist analysed
757 the samples and categorized them into tumor and matched normal breast tissue blocks.
758 Fresh tissue samples were transferred to sterile PBS buffer (Sangon Biotech, E607008)
759 on ice and processed immediately upon arrival at the laboratory. The breast cancer and
760 normal tissue samples were then dissociated, and viable single cells were isolated.
761 Single-cell whole genome amplification, library preparation, and whole-genome
762 sequencing were subsequently performed as described below.

763 **Isolation of single cells of breast tissue samples for scWGS**

764 Fresh tumor and matched normal breast tissue samples were collected and processed
765 within two hours post-surgery. For each sample, small tissue blocks of approximately
766 1 cm^3 were isolated for dissociation. The tissue dissociation procedure was performed
767 according to previously described methods[9,10], with slight modifications. Briefly, tissue
768 samples were manually minced in 1.5 mL tubes containing collagenase type 3 (300
769 U/mL) (Worthington Biochemical, #LS004182) and hyaluronidase (100 U/mL)
770 (Worthington Biochemical, #LS002592) for 2 hours, with gentle pipetting every 60
771 minutes. The digested tissue was then centrifuged at 1500 rpm for 5 minutes. The pellet
772 was digested with 5 mL of pre-warmed 0.25% trypsin-EDTA (Thermo #25200072) at
773 37°C for 5 minutes. Digested cells were washed once with HBSS + 2% FBS (Jackson,
774 #017-000-121), then treated with 0.1 mg/mL DNase I (Worthington Biochemical,
775 #LS002139) in DMEM/F12 media at 37°C for 5 minutes. After washing, the cells were
776 filtered through a 40 μm mesh. Dissociated cells were counted and resuspended in the
777 designated medium for subsequent staining and FACS sorting. Meanwhile, tissue
778 surrounding the small blocks (within 0.5 cm from the edge) was similarly dissociated
779 into a single-cell suspension. DNA was then extracted using a DNeasy Blood & Tissue
780 Kit (Qiagen, 69504) from these suspensions for bulk DNA whole genome sequencing.

781 For FACS sorting, single cells were stained with the following antibodies: CD45
782 (BioLegend, 304022), CD31 (BioLegend, 303114), CD49f (Thermo, 12-0495-82),
783 EpCAM (BioLegend, 324204), and CD34 (STEMCELL Technologies, 60013Az.1).
784 The antibodies were used at a dilution of 1:200, with appropriate conjugated
785 fluorophores, and incubated for 30 minutes on ice. After staining, cells were washed
786 and resuspended in Hanks Balanced Salt Solution (HBSS) + 2% FBS + Antibiotic-
787 Antimycotic + 4′,6-diamidino-2-phenylindole (DAPI) (1 μg/mL) at a density of 5
788 million cells/mL. The stained samples were then analyzed and sorted on a FACSAria
789 Fusion (BD Biosciences) with a 100-μm nozzle. The target single cells were sorted into
790 individual tubes, each containing a single cell and 3 μL of cell buffer, prepared for
791 subsequent single-cell whole genome amplification. The FACS data were further
792 processed using FlowJo V10. Gating strategy is shown in Extended Data Fig. 4.

793 **Whole genome amplification and single cell DNA & RNA sequencing**

794 The sorted single cells were subjected to whole-genome amplification using either
795 REPLI-G (Qiagen, 150023) or ResolveOME (Bioskyrb). Single cells amplified with
796 ResolveOME yielded both whole-genome amplification products and full-length
797 transcriptome amplification products. In some cases, both reactions were combined to
798 generate both whole-genome amplification products (via REPLI-G) and full-length
799 transcriptome amplification products from the same single cell. All whole-genome and
800 full-length transcriptome amplification products from ResolveOME underwent library
801 preparation using the ResolveOME library preparation protocol, including a 9-10 cycle
802 PCR library amplification. For REPLI-G whole-genome amplification products, as well
803 as the extracted paired bulk DNA, PCR-free library preparation was performed using
804 the Hieff NGS Ultima Pro PCR Free DNA Library Prep Kit V2 (Yeashen, 12196).

805 The WGS libraries were sequenced on the BGISEQ DNBSEQ-T7 platform (BGI Inc.)
806 using a paired-end sequencing length of 150 bp (PE150). The average sequencing depth
807 for each WGS library was 30–40x, while the bulk WGS library had an average
808 sequencing depth of 150x. Full-length transcriptome libraries were sequenced on the
809 NovaSeq X Plus platform (Illumina Inc.) using the PE150 strategy, generating
810 approximately 20 million reads per library. Additionally, parallel library preparation of
811 the same whole-genome amplification product or parallel sequencing of the same
812 library was conducted to obtain WGS data duplication.

813 **Processing of RNA sequencing data**

814 For RNA sequencing data, trimmomatic[11] (v 0.39) was first used to remove adapters.
815 Salmon[12] (v 1.10.2) was used to quantify the expression of transcripts. The salmon
816 index was constructed using the GTF and FASTA files (v GRCh37.87) from
817 ENSEMBL.

818 **Comparison of algorithms detecting mosaic STR mutations from the**
819 **benchmarking dataset**

820 To adapt the two germline mutation callers for identifying mosaic STRs from single
821 cell sequencing data, we applied a series of stringent filters to each software to generate
822 a mosaic mutation call set.

823 HipSTR[3] (v0.6.2) was run using the mode of "*de novo* stutter estimation + STR calling
824 with *de novo* allele generation", variant calls were then filtered following recommended
825 empirical cutoffs (--min-call-qual 0.9 --max-call-flank-indel 0.15 --max-call-stutter
826 0.15 --min-call-allele-bias -2 --min-call-strand-bias -2). ExpansionHunter[13] (v5.0.0)
827 with default settings were used for variant callings of benchmark datasets.

828 Germline genotypes and mosaic genotypes were identified as the most frequent and
829 second most frequent genotypes across all cells. Cell-specific variants were identified
830 by those that have one mutant cell, and shared variants were identified as those with
831 multiple mutant cells. In addition, if there are less than five valid cell calls, the candidate
832 loci were excluded from the mutation list.

For BayesMonSTR, variants were identified using a genotyping cutoff of mosaic posterior $\geqslant$ 0.9, predicted as mosaic by the RF model, and passing a series of hard filters. More specifically, for phasable sites, the hard filters include a series of phasing-related filters (equations 28-31, bulk discordant rate < 0.1, single cell discordant rate < 0.1, amplification error rate < 0.1 and mutant cell discordant rate < 0.1), and other filters (VAF $\geqslant$ 0.1 and mutant read count $\geqslant$2); For non-phasable sites, the hard filters include VAF > 0.25 and mutant read count $\geqslant$ 3.

Cross validation strategies of candidate mutations

To evaluate candidate STR loci detected by different callers, we assessed validation independently at three levels—DNA (duplicate library), RNA (single-cell RNA-sequencing), and bulk sample—using orthogonal evidence derived from the same single cell or its matched bulk sample. DNA-level validation was based on duplicate-library sequencing, RNA-level validation was based on single-cell RNA-sequencing reads, and bulk-level validation leveraged matched bulk sequencing from the same sample. Results from BayesMonSTR_phase, BayesMonSTR_unphase, HipSTR, and ExpansionHunter were first harmonized into a unified framework and then evaluated within each data_type × protocol_type stratum.

For DNA-level validation, a one-sided lower-tail binomial test was applied using the duplicate-library mutant read count (k) and total read count (n). For BayesMonSTR_phase and BayesMonSTR_unphase, two tests were performed, using either the model-derived expected mutant fraction (vaf_mut_sc_mean_prob) or $\theta = 0.5$ as the null probability, and the larger P value was retained for classification. For HipSTR and ExpansionHunter, the binomial test was performed using $\theta = 0.5$ only. For RNA-level validation, a one-sided lower-tail binomial test was applied using the RNA mutant read count (k) and total RNA read count (n), with $\theta = 0.5$ as the null probability. Loci with missing data or zero read depth were classified as undetermined. For the remaining loci, validation outcomes were defined as: validated when $P \geqslant 0.05$ and $k > 0$; unvalidated when $P < 0.05$ or $k = 0$; and undetermined otherwise.

For bulk-level validation, matched bulk sequencing data from the same biological sample were used to assess whether a candidate locus represented a somatic mutation or a germline variant. Two types of bulk data were utilized for this purpose: blood-derived bulk sequencing, serving as a constitutional germline reference, and normal tissue bulk sequencing from the same individual. A locus was classified as germline—and therefore assigned to the unvalidated category, as germline variants and sequencing artifacts do not constitute somatic mutations—if either of the following criteria was met in both the blood and normal bulk samples: (i) the variant allele fraction (VAF) in the bulk sample fell within the range of 0.4–0.6, consistent with heterozygous germline presence, or (ii) a two-tailed binomial test against the null hypothesis of $\theta = 0.5$ yielded $P > 0.05$, indicating that the allele frequency in bulk was statistically indistinguishable from a germline heterozygous state. Loci not meeting either germline criterion in both bulk samples were retained as candidate somatic mutations for downstream analysis.

When multiple callers reported the same locus within the same data type and amplification protocol but assigned discordant validation states, a predefined caller-priority rule was applied (BayesMonSTR_phase > BayesMonSTR_unphase > HipSTR > ExpansionHunter) to assign a final validation status. DNA-level, RNA-level, and bulk-level results were retained separately; no additional cross-layer priority rule was imposed among duplicate-library, RNA, and bulk evidence. Validation statistics were summarized using unique STR loci as the counting unit. Within each data_type × protocol_type × source stratum, we counted validated, unvalidated, and undetermined loci. The primary validation rate was defined as the number of validated loci divided by the total number of unique loci in that stratum, and, for display purposes, we additionally calculated the fraction of validated loci among clearly classified loci (validated + unvalidated).

Shared variants were further evaluated by manual inspection and PhyloSOLID (v 0.0.1) STR genome validation mode (https://github.com/douymLab/PhyloSOLID). Cell genotype likelihoods along with alternative and total read depth information were used for shared variants evaluation based on phylogenetic tree.

A gold-standard benchmarking dataset was constructed from 29 single cells derived from a breast tissue sample, comprising 7 epithelial cells from tumor tissue (Epi-T), 12 epithelial cells from normal tissue (Epi-N), and 10 immune cells (Imm). Of these, 12 cells had duplicate libraries (sequenced twice), enabling DNA-level validation, and 16 cells underwent both DNA and RNA sequencing from the same cell, enabling RNA-level validation. Single-cell DNA amplification was performed using either PTA or MDA. This dataset served exclusively as a benchmarking resource and is distinct from the public datasets used in cross-validation experiments. To prevent leakage between training and evaluation sets, cell-specific mutations derived from two cells within the benchmarking dataset were used as part of the model training set; these two cells were subsequently excluded from all test and validation analyses, ensuring an unbiased assessment of model performance on the remaining benchmarking samples.

To validate model performance on the public dataset (Supplementary Fig. 5g), we employed a leave-one-out cross-validation strategy. Specifically, when predicting and validating on a given dataset, the model was trained on all remaining datasets excluding the target dataset. For example, predictions for the MDA neuron dataset were generated using a model trained on all other datasets, and similarly, predictions for the single-cell colony skin dataset were generated using a model trained on all datasets except the single-cell colony skin data. This strategy ensures that the model has no prior exposure to the validation target during training, providing an unbiased assessment of generalization performance across independent datasets.

Visualization of STR mutations

913 INSIGHT (v 0.0.1) developed by our lab were used for visualization of STR mutations
914 (https://github.com/douymLab/INSIGHT).

915 **Identification of mosaic SNVs from single cells**

916 Similar strategies were applied to detect mosaic SNVs in the same datasets. After
917 sequence alignment and data preprocessing, raw candidate loci were identified by
918 mpileup in SAMtools[14] (v1.13) with the parameters '-B -d 8000 -q 10 -Q 10'. Then, the
919 mutations were identified by using MosaiSC-DNA[15] (v0.0.1,
920 https://github.com/douymLab/mosaicSC). The predicted SNVs were further filtered
921 with mutant cell VAF $\geqslant$ 0.3 and mutant read count $\geqslant$ 3 to get a more reliable set of
922 mosaic SNVs.

923 **Phylogeny reconstruction using shared variants**

924 Given the modest validation rate for shared variants, the constructed phylogenetic tree
925 (Figure 2g) was based on manually checked cell-by-mutation matrix (Supplementry
926 Table 5). PhyloSOLIDD[16] (v0.0.1) was used for phylogeny reconstruction.

927 **Correlation of age with multiple mutation-derived features**

928 We performed association analyses between age and multiple mutation-derived
929 features, including mutation rate, variant length, and cell proportion with >5bp indels.
930 Among these, the mutation rate was calculated by dividing the number of mutations by
931 the total length of callable regions. We used scipy[17] (v 1.14.1) to perform linear
932 regression analysis between age and features.

933 **Mutation signature analysis**

934 For mutations in STR regions, mutation matrices were generated using
935 SigProfilerMatrixGenerator[18] (v 1.2.31), and then, SigProfilerAssignment[19] (v 0.1.9)
936 was used to assign known COSMIC mutation signatures to every individual, which was
937 accomplished using the forward stagewise algorithm and nonnegative least squares. In
938 our samples, three 83-cat signatures of small insertions and deletions (ID1, ID2, ID12)
939 and multiple 96-cat signatures of single base substitutions were found to be active.
940 Additionally, SigProfilerAssignment was used to assign signature probabilities to every
941 mutation, allowing to quantify the mutation burden associated with each signature for
942 every individual.
943 Additionally, we also used a redefined InDel taxonomy[20] to illustrate mutation profile
944 for InDels in STR regions, and the mutation signature similarity was computed via
945 cosine similarity was used to compute similarity between our InDels and those caused
946 by the *MSH2* and *MSH3* knockouts[20].
947 We also compared the pattern of STR-region SNVs and the established patterns of
948 extra-STR mosaic SNVs in normal human prefrontal cortex[21] . The mutation signature
949 similarity was computed via cosine similarity.
950 Specially, for SNVs in STR regions, a custom method was developed to classify SNVs
951 into different categories. Initially, mutations were distinguished based on the six kinds
952 of fundamental base changes, followed by classification according to the nucleotide

953 context of the entire STR region in which the SNV occurred (considering both repeat
954 unit length and total length of the repeat region). Subsequently, non-negative matrix
955 factorization was performed using MutationalPatterns[22] (v 3.12.0) to extract three
956 distinct *de novo* mutation signatures. Based on the relative contribution of each
957 signature, the mutation burden associated with each signature for every individual was
958 quantified.

959 **Relationship between mosaic SNVs in DNA repair gene and mosaic STR indel**
960 **burden**

961 The list of DNA repair genes was obtained from the comprehensive compilation
962 maintained by the laboratory of Richard D. Wood at The University of Texas MD
963 Anderson Cancer Center (https://www.mdanderson.org/research/departments-labs-
964 institutes/labs/wood-laboratory/resources.html). This curated gene set represents an
965 updated synthesis based on previously published reviews.

966 ANNOVAR[23] (v 2020-06-07) was used to identify exonic variants, their associated
967 genes, and their effects on the protein sequence, including synonymous,
968 nonsynonymous, stop-gain, stop-loss, and frameshift variants. Variants annotated as
969 "unknown" by ANNOVAR due to incomplete ORF information were re-annotated
970 using the Ensembl Variant Effect Predictor[24] (VEP, https://www.ensembl.org/vep) to
971 determine their coding consequences based on updated transcript models.
972 To assess the association, we fitted a generalized linear mixed model (GLMM) with a
973 negative binomial distribution to account for overdispersion in count data. The model
974 included somatic SNV status (presence in DNA repair gene vs. absence) as a fixed
975 effect and a random intercept for individual to account for within-individual
976 correlations (e.g., multiple neurons or samples per individual). The significance of fixed
977 effects was evaluated using Wald tests. All analyses were performed in R (v 4.3.3)
978 using the glmmTMB[25] (v 1.1.14) package.

979 **Identification of potentially functional-disrupting mosaic STR mutations**

980 ANNOVAR was employed to determine the nearest gene of mosaic STR mutations and
981 identify potentially damaging exonic variants, including nonsynonymous, stopgain,
982 stoploss, and frameshift variants. CADD[26] (v 1.7) scores were incorporated to assess
983 variant deleteriousness. ENCODE's candidate cis-regulatory elements annotation[27] was
984 used to identify variants with potential regulatory functions. The chromatin state
985 annotation from ENCODE was used to identify variants in active TSS and enhancer
986 regions. Chromatin state annotations were derived from the ChromHMM 18-state
987 model from the ENCODE portal[28] for seven primary B cell samples, 17 lung tissues
988 and 16 brain tissues.

989 Potentially functional-disrupting STR indels were defined as variants in upstream, UTR
990 or exonic regions annotated by ANNOVAR, or regions defined as chromatin state of
991 active TSS and enhancer by ENCODE, or variants have CADD score greater than 20.

992 Gene affected by a STR indel was determined as the nearest gene annotated by
993 ANNOVAR.

994 **Identification of mosaic STR mutations in functional non-coding elements**

995 To identify STR mutations may have a significant impact on gene expression, we
996 utilized the human G-quadruplexes prediction panel from EndoQuad[29] for G-
997 quadruplex forming regions, the Level 2 R-loop annotation panel from R-loop Base[30]
998 for R-loop forming regions, ENCODE's candidate cis-regulatory elements
999 annotation[27] for CTCF-binding sites, and the CpG islands annotation from the UCSC
1000 Genome Browser[31] for CpG islands.

1001 **Identification of mosaic STR mutations near genes of high expression**

1002 Gene expression data of whole blood, lung and brain cortex tissues was downloaded
1003 from GTEx[32]. Within each tissue, we calculated quartiles of expression of all genes near
1004 STR loci and categorized each gene into four categories (Q1 to Q4) based on average
1005 expression level across samples, with Q1 representing expression level is in minimum
1006 to first quartile.

1007 **Gene ontology (GO) and disease enrichment**

1008 KOBAS[33] (v 3.0) was used to conduct Gene Ontology (GO) and Kyoto Encyclopedia
1009 of Genes and Genomes (KEGG) disease enrichment of genes near STR indels. To be
1010 specific, Fisher's exact test was used to calculate P-values, while Benjamini and
1011 Hochberg (1995) method was used to control the FDR.

1012 **Process of snATAC-seq data**

1013 Raw FASTQ files of snATAC-seq were aligned to the hg38 reference genome, using
1014 10x cellranger-atac count toolkit (v 2.1.0). Only cells identified through cellranger's
1015 cell calling were included in subsequent analyses. Duplicate reads were first filtered to
1016 ensure data integrity. Reads sharing identical cell barcodes and genomic coordinates
1017 (start and end positions) were grouped. Within each group, only reads with identical
1018 sequences were considered duplicates; among these, the read with the highest mapping
1019 quality and average base quality was retained.

1020 **Detection of mosaic STR mutations from snATAC-seq**

1021 Building upon the framework of BayesMonSTR, we adapted the approach to better
1022 handle the sparsity of single-cell ATAC-seq data by performing genotyping at the
1023 pseudo-bulk level.
1024 The stutter error model was trained using 26 snATAC-seq datasets derived from brain
1025 tissue and whole blood samples, obtained from the publicly available studies
1026 SRP309633 and SRP192525. After estimating the stutter error model and extracting
1027 alleles through STR segmentation as mentioned before, we directly estimated
1028 individual-level mosaic posterior probabilities at the pseudo-bulk level.
1029 We than perform multiple strict filters to retain highly confident mutations including
1030 1. posterior probability exceeding 0.5.

2. count of used reads in genotyping exceeding 10.
3. homozygous in the bulk genotype with exactly one alternative read.
4. no other reads at the locus contained off-target stutter artifacts or mismatches.
5. recurred in less than two individuals.

Per-cell mutation rate estimation

In snATAC-seq data, the mutation rate per cell can be estimated for each individual cell as the ratio of the number of detected mutations to the size of its callable region. Callable regions were defined as regions covered by at least one read spanning STR with high mapping quality. And the mutation rate of an individual (μ) can be estimated by the mean of cells' mutation rate:

$$\mu = \frac{\sum_{i=1}^{N_c}(\frac{M_i}{L_i})}{N_c} \tag{32}$$

Where:

- M_i: the number of somatic mutations detected in cell i;
- L_i: the total length (bp) of the callable genomic regions (i.e., sequenced with sufficient coverage and quality) in cell i;
- N_c: the total number of cells analyzed.

The mutation rate of the regulatory region was calculated using the same strategy, with the exception that regulatory callable regions were used instead of whole callable regions.

Dual-Luciferase promoter reporter assay

The promoter sequences of *NRN1* and *CNTNAP2* were PCR-amplified from HEK293T cDNA templates and cloned into our dual-luciferase reporter plasmid vector. All constructs were verified by sequencing. 0.2M HEK293T cells were seeded into 24-plate well with DMEM for 12 hours, cells were transiently transfected with plasmids using polyethyleneimine transfection reagent (1mg/mL; Polyscience, #23966-1). After 48 hours, promoter activity was analyzed using the Dual-Luciferase Reporter Gene Assay Kit (Yeasen Biotechnology, 11402ES60) following manufacturer's protocol on a Varioskan LUX Microplate reader (Thermo). Relative promoter activity was calculated as the ratio of firefly/Renilla luminescence.

1060 **References**

1061 1. Chen, S. *et al.* A genomic mutational constraint map using variation in 76,156 human
1062 genomes. *Nature* 1–11 (2023) doi:10.1038/s41586-023-06045-0.

1063 2. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel.
1064 *Nat. Genet.* **48**, 1443–1448 (2016).

1065 3. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat*
1066 *Methods* **14**, 590–592 (2017).

1067 4. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

1068 5. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular
1069 biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

1070 6. Dou, Y. *et al.* Accurate detection of mosaic variants in sequencing data without matched
1071 controls. *Nat Biotechnol* **38**, 314–319 (2020).

1072 7. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI
1073 for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).

1074 8. Ewing, A. D. *et al.* Combining tumor genome simulation with crowdsourcing to benchmark
1075 somatic single-nucleotide-variant detection. *Nat Methods* **12**, 623–630 (2015).

1076 9. Yuan, L. *et al.* Reconstruction of dynamic mammary mini gland in vitro for normal
1077 physiology and oncogenesis. *Nat. Methods* **20**, 2021–2033 (2023).

1078 10. Fu, A. *et al.* Tumor-resident intracellular microbiota promotes metastatic colonization in
1079 breast cancer. *Cell* **185**, 1356-1372.e26 (2022).

1080 11. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
1081 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

1082 12. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and
1083 bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

1084 13. Dolzhenko, E. *et al.* ExpansionHunter: A sequence-graph based tool to analyze variation in
1085 short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).

1086 14. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
1087 2079 (2009).

1088 15. Zheng, Y., Fan, W., Lu, J., Yang, Q. & Dou, Y. Accurate detection of specific and shared
1089 mosaic variants in single-cell whole-genome sequencing data in normal human cells. (2025).

1090 16. Yang, Q. *et al.* PhyloSOLID: robust phylogeny reconstruction from single-cell data despite
1091 pervasive errors and extreme sparsity.

1092 17. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python.
1093 *Nat. Methods* **17**, 261–272 (2020).

1094 18. Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring
1095 patterns of small mutational events. *BMC Genom.* **20**, 685 (2019).

1096 19. Díaz-Gay, M. *et al.* Assigning mutational signatures to individual samples and individual
1097 somatic mutations with SigProfilerAssignment. *Bioinformatics* **39**, btad756 (2023).

1098 20. Koh, G. C. C. *et al.* A redefined InDel taxonomy provides insights into mutational
1099 signatures. *Nat. Genet.* **57**, 1132–1141 (2025).

1100 21. Miller, M. B. *et al.* Somatic genomic changes in single Alzheimer's disease neurons. *Nature*
1101 1–9 (2022) doi:10.1038/s41586-022-04640-1.

1102 22. Manders, F. *et al.* MutationalPatterns: the one stop shop for the analysis of mutational
1103 processes. *BMC Genom.* **23**, 134 (2022).

1104 23. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants
1105 from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).

1106 24. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

1107 25. Brooks, M., E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-
1108 inflated Generalized Linear Mixed Modeling. *R J.* **9**, 378 (2017).

1109 26. Schubach, M., Maass, T., Nazaretyan, L., Röner, S. & Kircher, M. CADD v1.7: using
1110 protein language models, regulatory CNNs and other nucleotide-level scores to improve
1111 genome-wide variant predictions. *Nucleic Acids Res.* **52**, D1143–D1154 (2024).

1112 27. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse
1113 genomes. *Nature* **583**, 699–710 (2020).

1114 28. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data
1115 portal. *Nucleic Acids Res.* **48**, D882–D889 (2019).

1116 29. Qian, S. H. *et al.* EndoQuad: a comprehensive genome-wide experimentally validated
1117 endogenous G-quadruplex database. *Nucleic Acids Res.* **52**, D72–D80 (2023).

1118 30. Lin, R. *et al.* R-loopBase: a knowledgebase for genome-wide R-loop formation and
1119 regulation. *Nucleic Acids Res.* **50**, D303–D315 (2021).

1120 31. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* **53**,
1121 D1243–D1249 (2024).

1122 32. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–
1123 585 (2013).

1124 33. Bu, D. *et al.* KOBAS-i: intelligent prioritization and exploratory visualization of biological
1125 functions for gene enrichment analysis. *Nucleic Acids Res.* **49**, W317–W325 (2021).

1126

Data availability

Bulk WGS data and single cell data from breast cancer samples generated in this study will be publicly available in the Sequence Read Archive (SRA) (PRJNA1378499). WGS data for neurons, B cells, and lung epithelium are accessible through dbGaP under accession numbers phs001485.v3.p1 and phs001808.v1.p1, and in the European Genome-phenome Archive (EGA) under accession EGAD00001005193, respectively. sn-ATAC-seq data from brain tissue are available in the EGA (accession EGAS00001001934) and Synapse (accession syn52293417).

Code availability

BayesMonSTR is implemented in Python and R and is licensed under the MIT License. The source code, documentation and examples are available on GitHub at https://github.com/douymLab/BayesMonSTR.

Acknowledgements

We are deeply grateful to the generous tissue contributions. We are grateful to Dr. Peter J. Park for his critical insights and discussions; Drs. Xushen Xiong, Manolis Kellis, Xiaodong Liu, Jing Zhang, Shumin Duan and Dan Zhou for data sharing and discussions; and Drs. Shang Cai, Danyang He, Hongtao Yu, Xiaochun Yu, Wei Wu, Lovelace J. Luquette, Ge Gao, Kai Zhang, Yanxiao Zhang, Jia Zheng, Yuxiang Huang, and Dangsheng Li for their comments and assistance. This work was supported by the Westlake Laboratory of Life Sciences and Biomedicine (Hangzhou 310024, Zhejiang, China), under the grant "Key R&D Program of Zhejiang" (2024SSYS0032) as well as the National Natural Science Foundation of China (32270682) to Y.D.; and the National Natural Science Foundation of China (82525055 and 82473121) to Q.X. We thank the Flow Cytometry Core Facility and the Genomics Core Facility at Westlake University for their experimental assistance, and the High-Performance Computing Center for technical support. We would like to thank the submitters of the following datasets: dbGaP accession phs001808.v1.p1, dbGaP accession phs001485.v3.p1, the European Genome-phenome Archive (EGA) accession EGAD00001005193, the EGA accession EGAS00001001934 and Synapse accession syn52293417. Data used in this study under the dbGaP accession phs001485.v3.p1 was generated by Drs. Christopher A. Walsh and Peter J. Park with funding from NINDS grant R01NS032457. The public datasets were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap and the EGA at https://ega-archive.org/.

Author contribution

Y.D. conceived and supervised the project and secured the funding. Q.X., D.H., X.L., S.C., X.X., M.K. provided support in the supervision of the project. C.W., W.F., W.W., Y.X., J.L., X.M. and J.Y. all made significant contributions to the project. W.F. contributed to software development with help from C.W., W.W., Y.X., J.L. and Y.Z., C.W. developed the pipeline analyzing snATACseq data and lead the bioinformatics analysis. Y.X contributed to phylogeny reconstruction, public data downloading and

developed the mutation-visulization software. Q.Y. helped phylogeny reconstruction. J.L. was responsible for generating the benchmarking dataset with help from X.Y., X.M., Y.L., and Z.M., X.Y. performed the breast tumor surgery. J.Y. and X.M. performed the functional assays. N.Y. contributed to cell sorting and related wet lab experiments. Y.Z. and J.L. were responsible for generating non-STR SNV mutation callings. W.L, W.W., C.W., Y.X, W.F. and H.L. contributed to mutation examination. Y.D., C.W., W.F., W.W. and J.L. wrote the manuscript. All authors carefully reviewed and approved the final manuscript. All authors in the Data Group contributed to data access, data generation, or data transfer.

Inclusion & Ethics statement

This study was conducted in accordance with ethical guidelines and principles. It involved one human participant, and no animal subjects were included. The authors affirm that all research procedures were performed with respect for the participant's rights and well-being. The sequencing data generated in this study are publicly available, and no personally identifiable information was included. The authors ensure that the study adheres to the highest ethical standards in research, data generation, and usage.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work the authors used DeepSeek in order to polish our language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Competing interests

The authors declare no competing interests.